Erschienen in: Dalmas, Martine/Fabricius-Hansen, Cathrine/Schwinn, Horst (Hrsg.): Variation im europäischen Kontrast. Untersuchungen zum Satzanfang im Deutschen, Französischen, Norwegischen, Polnischen und Ungarischen. Berlin/Boston: de Gruyter, 2016. (Konvergenz und Divergenz 5), S. 9-52.

Hagen Augustin

Quantitative Untersuchungen zum deutschen Vorfeld und seinen Äquivalenten in sechs verschiedensprachigen Wikipedia-Korpora

Abstract: Vorgestellt werden kontrastive Analysen zur Besetzung und Häufigkeitsverteilung von Vorfeldern im Deutschen und ihren französischen, italienischen, norwegischen, polnischen und ungarischen Äquivalenten in morphosyntaktisch annotierten Wikipedia-Korpora. Im Rahmen der Untersuchung wurden mit korpusanalytischen Methoden quantitative Zusammenhänge bei den sprachspezifischen Ausprägungen von Vorfeldern nachgewiesen, die im Einklang mit typischen Struktureigenschaften der untersuchten Kontrastsprachen stehen. Die Ergebnisse legen aber nahe, dass die untersuchten Vorfeldstrukturen – trotz der beträchtlichen Größe und thematischen Vielfalt der Wikipedia-Korpora – nicht hinreichend repräsentativ sind, um uneingeschränkt Rückschlüsse auf allgemeine Struktureigenschaften der sechs Kontrastsprachen zu ziehen. Hierfür verantwortlich ist insbesondere die ausgeprägte Textsortenspezifizität der Mediengattung (Online-)Enzyklopädie, was mithilfe weiterer Vergleichskorpora aufgezeigt werden konnte.

1 Einleitung/Thematik

Der vorliegende Aufsatz befasst sich sowohl mit der methodischen Konzeption als auch den Resultaten von kontrastiven Analysen[1] deutsch-, französisch-, italienisch-, polnisch- und ungarischsprachiger Wikipedia-Korpora, deren Ziel es ist, Umfang und Häufigkeitsverteilung von Wortmaterial im Satzbereich vor dem finiten Verb quantitativ zu erfassen und zu dokumentieren. Im deutschen Stellungsfeldermodel ist dieser topologische Bereich teilidentisch mit dem Vorfeld, das

1 Die Arbeiten wurden im Rahmen des europäischen Kooperationsprojektes EuroGr@mm am Institut für Deutsche Sprache und an den Partneruniversitäten im Ausland durchgeführt. Schwerpunkt des Forschungsprojekts bildete für den Zeitraum von 2010 bis 2012 die korpusgestützte kontrastiv vergleichende Erforschung der grammatischen Variation im standardnahen Deutsch, wobei insbesondere der topologische Satzbereich vor dem finiten Verb in den Vergleichssprachen unter morphosyntaktischen und informationsstrukturellen Gesichtspunkten im Fokus des Interesses stand.

nach spezifischen Regeln (Zifonun/Hoffmann/Strecker 1997: 1576ff.) besetzt wird. Der Fokus der Untersuchung liegt dabei auf der textsortenspezifischen Ausprägung dieser, in den einzelnen Kontrastsprachen sehr unterschiedlichen Besetzungsmöglichkeiten.

2 Wikipedia-Artikel als Textsorte

Die empirische Grundlage der hier vorgestellten Untersuchungen bilden sechs Volltextkorpora, die den Bearbeitungstand des Jahres 2011 der deutschen (DE), französischen (FR), italienischen (IT), norwegischen (NO)[2], polnischen (PO) und ungarischen (UN) Sprachversion der Online-Enzyklopädie Wikipedia wiedergeben. Einen Überblick über die Entstehung und technische Aufbereitung der Korpora geben Bubenhofer/Haupt/Schwinn (2011). Bei den Wikipedia-Artikeln handelt es sich um freie Inhalte, was sich aus urheberrechtlichen Gründen bei der Korpuserstellung als Vorteil erweist. Bei der Korpusanalyse wurden Textsortenspezifika, insbesondere die Textstruktur und die sprachlichen Merkmale der Wikipedia-Artikel berücksichtigt. Mit einer textsortenspezifischen Einordnung von Wikipedia-Artikeln haben sich Fandrych/Thurmair (2011) beschäftigt, die sie als eine „vom Prototyp etwas abweichende Variante der Textsorte ‚Lexikonartikel'" (ebd.: 104) betrachten. Wikipedia unterscheidet sich von herkömmlichen Enzyklopädien u.a. dadurch, dass jeder aktive Benutzer gleichzeitig als Autor und Redakteur tätig werden kann, an einem Wikipedia-Artikel also eine Vielzahl von Personen mitgewirkt haben können. Die i.d.R. anonym ausgeübte Autorenschaft einzelner Texte bzw. Textpassagen kann nur indirekt über die Versionsgeschichte eines Artikels erschlossen werden und wurde bei der Erstellung der Korpora nicht erfasst.

Die sprachliche Gestaltung eines aktuellen Wikipedia-Artikels kann also nicht individuellen Autoren zugeordnet werden, obliegt aber potenziell einer indirekten Kontrolle durch alle Benutzer, die Artikel auch verändern können. Auf eine Reihe spezifischer Eigenheiten von Wikipedia-Texten, die sich in sprachlichen Merkmalen niederschlagen und in der Korpusanalyse berücksichtigt werden müssen, wird im Folgenden eingegangen.

2 Es wurde nur die Version „Norsk bokmål" berücksichtigt.

2.1 Äußere und innere Artikelstruktur

Wikipedia verfügt über keine redaktionell bearbeitete Stichwortliste. Die thematische Breite der Inhalte ist enorm, aber z.T. quantitativ und qualitativ unausgewogen. Einzelne Themengebiete werden oft durch identisch aufgebaute, in Serie verfasste Artikel konstituiert (z.B. zu Asteroiden, Gemeinden, Tier- und Pflanzenspezies), die nicht selten identische bzw. teilidentische Sätze enthalten, bei denen sich ggf. nur das Subjekt und die Numerialia unterscheiden. Es handelt sich bei solchen Serienartikeln also um Quasi-Dubletten. Für zahlreiche Themengebiete bietet das Wikipedia-Autorenportal zudem fertige, die Textstruktur vordefinierende Formatvorlagen an, die beim Erstellen eines neuen Artikels verwendet werden können.

Die allgemeine Textstruktur von Wikipedia-Artikeln ähnelt der von gedruckten Lexikonartikeln (Fandrych/Thurmair 2011: 107), d.h., die Texteinheiten der Artikel sind nach rekurrenten Prinzipien strukturiert und sprachlich ausgestaltet. Das Denotat des Stichworts (in Abhängigkeit der Gegenstandsklasse, z.B. Menschen, Pflanzen, Städte, historische Ereignisse usw.) bestimmt dabei maßgeblich die Struktur und die sprachlichen Merkmale des Artikels (Hoffmann 1988: 154ff.). Die Grundinformation über ein Stichwort wird häufig über die „klassische, einordnende Art der Definition" vermittelt (Fandrych/Thurmair 2011: 96), die mit Hilfe dafür typischer sprachlicher Mittel realisiert wird. Dabei weisen enzyklopädische Werke die Tendenz auf, Substantive als Stichwörter zu bevorzugen (ebd.: 96). Mit Inhaltsverzeichnissen, Zwischenüberschriften, Tabellen, Aufzählungen, Medien-Inhalten u.Ä. setzen Wikipedia-Artikel sehr häufig textstrukturierende Elemente mit besonderen sprachlichen Merkmalen ein, die in dieser Form in gedruckten Lexikonartikeln nur spärlich oder gar nicht vorkommen. Sie bedürfen, sofern sie nicht entfernt wurden, einer besonderen Annotation bei der Korpuserstellung bzw. -analyse, die sie vom eigentlichen Textkörper unterscheidbar machen.

2.2 Stilistik und grammatische Strukturmerkmale

Fandrych/Thurmair (2011: 108–113) heben bei deutschsprachigen Lexikonartikeln folgende grammatische Strukturmerkmale als textsortenspezifisch hervor:

- Bevorzugung des Nominalstils
- häufige Verwendung von Parenthesen und Ellipsen (v.a. der Kopula)
- asyndetischer Stil (wenig Konnektoren)
- hohe Frequenz des Passivs
- geringe Frequenz von Pronomina

Die stilistischen Besonderheiten betrachten sie v.a. unter dem Gesichtspunkt der sprachlichen Verdichtung und weisen in diesem Zusammenhang auf erkennbare Unterschiede zwischen traditionellen, gedruckten Lexikonartikeln und Wikipedia-Artikeln hin, die länger/ausführlicher und sprachlich weniger nominal verdichtet (ebd.: 107) sind. Dieser Befund muss im Lichte konkreter Vorgaben betrachtet werden, die Wikipedia-Autoren zum Verfassen von Artikeln bekommen. Ihnen werden zu allen Beteiligungsformen, darunter auch zu Textstruktur, Stil und Typografie der Artikel, umfassende Richtlinien an die Hand gegeben[3], die auf den Wikipedia-Hilfeseiten für alle sechs Kontrastsprachen nachzulesen sind. Die Art der Hinweise unterscheidet sich allerdings in den unterschiedlichen Sprachversionen zum Teil erheblich, nicht zuletzt auch aufgrund kultureller Unterschiede und sprachspezifischer Charakteristiken der Textsorte Lexikonartikel. Welchen Einfluss stilistische Vorgaben auf die grammatischen Strukturmerkmale der Texte haben können, soll exemplarisch an der deutschsprachigen Wikipedia gezeigt werden.

Um dem Anspruch gerecht zu werden, auch für Laien problemlos verständlich zu sein, beruft sich die Wikipedia-Seite „Allgemeinverständlichkeit" im Abschnitt „Gutes Deutsch" auf Zimmer (2007), gespickt mit einigen sehr plakativen Ratschlägen und Kommentaren (in Klammern), vgl. Wikipedia (2014a):

- Satzlänge („Kurze Sätze sind besser als lange. Keep it simple.")
- Satzgliedstellung („Subjekt–Prädikat–Objekt ist nie verkehrt.")
- Subordination („Höchstens ein Nebensatz."; „Nebensatz vor oder nach dem Hauptsatz. Schachtelsätze vermeiden."; „Thomas Mann wäre kein guter Wikipedia-Autor.")
- Gebrauch von Konnektoren („Nachgestellte Kausalsätze mit *weil*, vorangestellte mit *da* beginnen."; „Neuer Hauptsatz mit *denn* ist klar und zwingt zu Kürze."; „Vor *denn* und *aber* Semikolon.")
- Vermeidung von Nominalstil („Substantive sind schwach, Verben sind stark.")
- Vermeidung des Passivs („Passivsätze möglichst vermeiden."; „Partizipialsätze sparen (passive) Verben.")

3 Da es keine systematische redaktionelle Bearbeitung gibt, die eine Umsetzung solcher Richtlinien konsequent gewährleistet, ist auf diesem Gebiet mit Varianz zu rechnen, die ggf. z.B. durch gezielte stilometrische Untersuchungen empirisch nachzuweisen wäre. Im Rahmen der vorliegenden Untersuchung konnte dieser Varianztyp aber nicht berücksichtigt werden.

Die Hilfeseite „Wie schreibe ich gute Artikel" (Wikipedia 2014c) nennt unter der Überschrift „Stil" Maßgaben, die sich inhaltlich mit den oben genannten Punkten weitgehend decken, aber differenzierter erläutert werden. Wikipedia-Artikel sollen demnach „in Stil und Ausdruck nicht umgangssprachlich verfasst sein, sondern in standardisierter Schriftsprache, [...]". Die Maßgaben richten sich aber insbesondere gegen zu starke inhaltliche und sprachliche Verdichtung und plädieren für einen einfachen Satzbau (z.B. „Verben nach vorne" – eine Maßgabe, die sich unmittelbar auf den hier untersuchten topologischen Satzbereich im Deutschen auswirken kann). Des Weiteren werden Richtlinien gegeben, die Tempus (Vermeidung des historischen Präsens) und Temporaldeixis (Vermeidung von relationalen Temporaladverbialia wie „im vorigen Jahr, letzte Woche, vor kurzem, derzeit, neuerdings, heute") betreffen. Auch zu Datumskonventionen und Typografie werden Hinweise gegeben. Universelle, textsortenabhängige Richtlinien, wie z.B. die Vermeidung relationaler Temporaladverbiala zugunsten von konkreten Zeitangaben, finden sich in fast allen Sprachversionen. In anderen Fällen gibt es unterschiedliche Bewertungen, z.B. empfehlen die ungarischen Richtlinien (Wikipedia 2014d) explizit die Verwendung des historischen Präsens und schreiben auch ausdrücklich die Kopula-Ellipse in einleitenden Sätzen vor. Die polnischen Richtlinien (Wikipedia 2014e) dulden generell Ellipsen in einleitenden Sätzen, betonen aber, dass sie im übrigen Text zu vermeiden seien. Im Unterschied zum Deutschen, Französischen, Italienischen und Norwegischen ist die Kopula-Ellipse im Polnischen eine grammatisch lizenzierte Erscheinung, die nicht ausschließlich ein textsortenspezifisches Merkmal darstellt.

Die deutschen Richtlinien sind in einigen Punkten, z.B. „Schreibe in ganzen Sätzen", als bewusste Distanzierung zum Stil, der in herkömmlichen (gedruckten) enzyklopädischen Lexika gepflegt wird, zu betrachten. Die Notwendigkeit der sprachlichen Verdichtung bei Büchern aufgrund des begrenzten Umfangs ist medial bedingt. Internet-Publikationen sind diesbezüglich weniger eingeschränkt. Ein exemplarischer Vergleich zwischen dem Stichwort „Mannheim" in der (gedruckten) Brockhaus-Enzyklopädie von 2006 und dem deutschen Wikipedia-Korpus von 2011 veranschaulicht, welche Unterschiede für die jeweilige formale Ausprägung der Textsorte charakteristisch sind:

(1) Mannheim, Stadt in Bad.-Württ., Stadtkreis im Reg.-Bez. Karlsruhe, mit 307500 Ew. zweitgrößte Stadt von Bad.-Württ., liegt in der Oberrheinebene an der Mündung des kanalisierten Neckars in den Rhein, 97 m. ü. M., bildet mit dem auf der anderen Rheinseite gelegenen Ludwigshafen (zwei Brücken) das Zentrum des Ballungsraumes Rhein-Neckar. [...] (Brockhaus-Enzyklopädie 2006: 622)

(2) <Mannheim> Die Universitätsstadt Mannheim ist mit etwa 315.000 Einwohnern die zweitgrößte Stadt Baden-Württembergs. Die ehemalige Residenzstadt (1720–1778) der historischen Kurpfalz bildet das wirtschaftliche und kulturelle Zentrum der europäischen Metropolregion Rhein-Neckar. Von seiner rheinland-pfälzischen Schwesterstadt Ludwigshafen am Rhein (164.000 Einwohner) ist Mannheim durch den Rhein getrennt. (http://de.wikipedia.org/wiki/Mannheim, Stand 2011)

Beide Texte weisen typische Verdichtungsmerkmale auf, die sich in der relativen Häufigkeit von komplexer Attribuierung und Erweiterungsnomina zeigen. Anaphorische Pronomina kommen nicht vor. Ihr Einsatz würde zwar zur Verdichtung beitragen, entspricht aber stilistisch nicht der Textsorte, die andere Mittel zur Herstellung von Kohärenz im Text bevorzugt (Hoffmann 1988: 142, 155f.). Beleg (1) verdichtet durch Hinzufügung von Informationen, die als Apposition zum Stichwort Mannheim oder als asyndetischer Teilsatz mit Kopulaellipse verstanden werden können. Der Beleg enthält viele Abkürzungen. Junktoren bzw. Pronomina, die zur Wiederaufnahme des Subjekts vonnöten wären, werden ausgelassen. Beleg (2) ist in vielerlei Hinsicht repräsentativ für die von Fandrych/Thurmair (2011) konstatierten sprachlichen Abweichungen der Wikipedia-Artikel vom Prototyp des Lexikonartikels. Beleg (2) ist im Vergleich zu (1) viel weniger stark verdichtet. Es handelt sich um vollständige Sätze ohne Ellipsen. Verdichtung findet vergleichsweise behutsam statt (Komposita, Genitivattribute, komplexe Adjektive, Parenthesen, Passivsatz). Beide Belege haben eine vergleichbare Anzahl von Wörtern, der Wikipedia-Beleg verzichtet jedoch auf Abkürzungen. Der Informationsgehalt der beiden einleitenden Absätze ist quantitativ ungefähr gleich, wobei inhaltlich unterschiedliche Akzente gesetzt werden. Die aktuelle Fassung[4] des einleitenden Absatzes des Wikipedia-Artikels „Mannheim" enthält aber mittlerweile mehrere Ergänzungen (und keine Kürzungen) und ist um über 30 Wörter angewachsen. Der gedruckte Brockhaus-Artikel ist insgesamt deutlich kürzer als der Wikipedia-Artikel.

3 Wikipedia-Korpora

Die Quelltexte der sechs untersuchten Wikipedia-Sprachversionen wurden mit Hilfe eines eigens zu diesem Zweck entwickelten Verfahrens (Bubenhofer/Haupt/

4 Mit dem Bearbeitungsstand vom 22. Juli 2014, 23:07 UTC.

Schwinn 2011) konvertiert. Die Größe (in Millionen Wörtern: MW) der Korpora ist sehr unterschiedlich (vgl. Tabelle 1). Sie ergibt sich aus dem Umfang der jeweiligen sprachspezifischen Wikipedia-Ausgaben zum Zeitpunkt der Korpuserstellung und aus dem Umstand, dass im Falle des ungarischsprachigen Korpus nur eine Stichprobe von Wikipedia-Artikeln zur Verfügung stand.

Tab. 1: Größe der Wikipedia-Korpora

Sprache	Größe (MW)
Deutsch	551,09
Französisch	527,94
Italienisch	329,06
Norwegisch (Bokmål)	70,38
Polnisch	190,05
Ungarisch (Stichprobe, Version CQPweb))	10,86

Die morphosyntaktische Annotation wurde mit Hilfe von verschiedenen, auf die einzelnen Kontrastsprachen abgestimmten Part-of-Speech-Taggern durchgeführt (hierzu ausführlich: Bubenhofer/Haupt/Schwinn 2011: 142). Recherchiert wurde mit dem Korpusrecherche- und -analysesystem COSMAS II (Institut für Deutsche Sprache 1991–2010)[5] und im Falle der ungarischsprachigen Wikipedia mit der Abfragesprache CQPweb (s. 3.3).

3.1 Vorfeld vs. linkes Feld

Ausgangspunkt ist die für das Deutsche etablierte, auf Drach (1937) zurückgehende Felderterminologie, die den Satz ausgehend von den Verbpositionen in Stellungsfelder einteilt. Ein Charakteristikum der deutschen Wortstellung ist, dass finite Hilfs- oder Modalverben in Verberst- und Verbzweitsätzen oft durch andere Satzelemente vom Rest des Verbalkomplexes getrennt sind. Die topologischen Felder werden durch Satzklammerteile voneinander abgegrenzt, wobei der linke Satzklammerteil bei Verberst- und Verbzweitsätzen durch die finite Verbform realisiert wird (Zifonun/Hoffmann/Strecker 1997: 1500). Im Rahmen des topologischen Feldermodells beschreibt das Vorfeld einen Teil des Satzes, der sich vor dem linken Satzklammerteil befindet. Es ist nur in Verbzweitsätzen

5 Das deutschsprachige Wikipedia-Korpus ist als Teilkorpus in DeReKo integriert (Korpussiglen WPD11) und öffentlich zugänglich.

besetzt, d.h. Verberst- und Verbletztsätze verfügen über kein Vorfeld. Dies bedeutet aber mitnichten, dass in diesen Satztypen dem linken Satzklammerteil keinerlei Stellungseinheiten vorausgehen können. In allen drei verbstellungsbasierten Satztypen können auch Einheiten diesen Bereich besetzen, die nicht als Stellungseinheiten des Vorfelds gelten, da sie nicht in der gleichen Weise wie Vorfeldeinheiten am syntaktischen Aufbau des Satzes beteiligt sind (Zifonun/Hoffmann/Strecker 1997: 1577) und als Einheiten des linken Außenfelds (ebd.) bezeichnet werden. Die für diese Position infrage kommenden Stellungseinheiten sind von ganz unterschiedlichem Aufbau und Umfang. Altmann bezeichnet solche Sachverhalte als „Störfaktoren" (Altmann 1987: 33) bei der Bestimmung der Verbstellung und nennt einige Beispiele: koordinative Konjunktionen („Aber läßt du das Buch da?"), Herausstellungsstrukturen und verlagerte Teilstrukturen[6] („Den neuen großen BMW – hast du den schon gesehen?" „Die Brigitte – die kann ich schon gar nicht leiden."). Darüber hinaus wären noch die bei Altmann (1987) nicht erwähnten „interaktiven Einheiten" (Zifonun/Hoffmann/Strecker 1997: 1577–1578) wie Interjektionen, Responsive und Anredeformen zu nennen, wie man sie in den Wikipedia-Korpora vor allem bei Zitaten, Redewiedergabe oder bei der Kommunikation in Diskussionsforen findet (3). Als „störend" bei der Verbstellungsbestimmung erweisen sich auch Vorfeldellipsen, wobei für Satzanfänge v.a. die von Altmann (1987: 33) genannten Beispiele Witz und Telegrammstil relevant sind (6).

Die folgenden Korpusbelege sollen einen Eindruck von generellen, nicht korpusspezifischen Besetzungsmöglichkeiten des Bereichs links vom ersten finiten Verb eines Satzes im Deutschen geben. Er kann z.B. Außenfeld- mit Vorfeldeinheiten kombinieren (3) bzw. ausschließlich Außenfeldeinheiten beinhalten (4). Er kann auch mit syntaktisch komplexen Vorfeldeinheiten besetzt werden (5) und sogar in Verbzweitsätzen unbesetzt bleiben (6). Diese Satzbereiche und die unmittelbar nachfolgenden finiten Verben werden jeweils in eckigen Klammern angegeben, die hier angenommenen Satzgrenzen ergeben sich aus der (interpunktionsbasierten) Korpus-Annotation:

(3) [„Nein, Herr Doktor, unsere Wege auf dem dichterischen und volkstümlichen Gebiete] [gehen] weit auseinander, ebensoweit wie unsere Dialekte." (http://de.wikipedia.org/wiki/Klaus_Groth, Stand 2011)

(4) [Oder] [hat] der Verleger den Notentext korrumpiert? (http://de.wikipedia.org/wiki/Urtext, Stand 2011)

6 Was (Altmann 1987) formal unter diesen „Teilstrukturen" im Einzelnen versteht, bleibt allerdings unklar.

(5) [[Dass ihm das zum Verhängnis werden] [würde,]] [ahnte] er nicht. (http://de.wikipedia.org/wiki/E._T._A._Hoffmann, Stand 2011)

(6) „Meine Lieben. [Ø] [Bin] gut gelandet, es geht gut. [Ø] [Komme] nach Oberschlesien, noch in Deutschland. Herzliche Grüße und Küsse euer Juller." (http://de.wikipedia.org/wiki/Julius_Hirsch, Stand 2011)

Dem (in der Linearstruktur des Satzes) ersten finiten Verb können sog. linke Außenfeldeinheiten (Zifonun/Hoffmann/Strecker 1997: 1577) vorausgehen (3, 4) – im Falle von linksversetzten Verbletztsätzen sogar ein ganzes (Nebensatz-)Mittelfeld + linker Satzklammerteil (5). In (4) liegt ein Verberstsatz vor, dem finiten Verb geht ein Konjunktor voraus. Beispiel (6) ist strukturell ein Verbzweitsatz, bei dem das finite Verb aber satzinitial steht, da eine Vorfeldellipse vorliegt. Die Beispiele deuten an, dass eine topologische Analyse der Korpusdaten, die allein auf der Verbstellung beruht, nicht hinreichend zuverlässig ist, um für das Deutsche Vorfeldeinheiten im engeren Sinne zu bestimmen. Gleichzeitig wirft die Variationsbreite von Stellungseinheiten, die im Deutschen dem finiten Verb vorausgehen können, die Frage auf, welchen konzeptionellen Zuschnitt ein für eine kontrastiv angelegte Studie geeignetes Tertium Comparationis aus dem Bereich der Wortstellung haben muss.

Da eine auf den Textbereich vor dem ersten finiten Verb eines Satzes maßgeschneiderte Suchanfrage in deutschsprachigen Korpora nicht nur Vorfeldeinheiten von Verbzweitsätzen, sondern auch andere Stellungseinheiten liefert, kann dieser empirisch ermittelte Textbereich also nicht einfach mit dem Vorfeld gleichgesetzt werden. Aus diesem Grund wurde im Rahmen der vorliegenden Untersuchung die Bezeichnung „linkes Feld" gewählt. Die Besetzungsmöglichkeit der Stellungsfelder und die Abfolge der Stellungseinheiten ist aber nicht nur im Deutschen an ganz bestimmte Regularitäten geknüpft, die auf unterschiedlichen Ebenen wie z.B. der Syntax oder der Informationsstruktur operieren. In anderen Sprachen unterliegt die Unterteilung von Sätzen in bestimmte, nach linguistischen Kriterien definierte Bereiche ganz anderen Voraussetzungen. Auch aus diesem Grund wurde in dieser Untersuchung als Tertium Comparationis der allgemein gefasste Begriff des linken Feldes verwendet. Dieses linke Feld wird ausdrücklich nicht als Stellungsfeld (mit Stellungseinheiten als dessen Elemente) im Sinne einer sprachspezifischen Topologie betrachtet, sondern soll lediglich für die korpusbasierte sprachübergreifende Erfassung des diskutierten Satzbereichs als Bezeichnung für das Wortmaterial dienen, das dem finiten Verb in den Kontrastsprachen vorausgeht. Es soll ausdrücklich nicht unterstellt werden, dass die Kontrastsprachen Struktureigenschaften besitzen, die von einem linken Feld als Stellungsfeld mit spezifischen Besetzungsregularitäten ausgehen. Die Frage,

ob es z.B. eine Grammatik des linken Felds im Polnischen usw. gibt oder nicht, kann im Rahmen der vorliegenden Analyse nicht beantwortet werden.

3.2 Suchmethode in den Korpora

Ungeachtet der sich abzeichnenden Probleme bei der Verbstellungsbestimmung wurden die Textbereiche zwischen einer Satzgrenze und der ersten nachfolgenden finiten Verbform zunächst mithilfe einer entsprechenden Suchanfrage in den hier untersuchten, nach Wortklassen (Parts-of-Speech) annotierten Wikipedia-Korpora gesucht. X und Y bezeichnen hier einzelne Wortformen (Tokens) im Korpus, nicht Stellungseinheiten:

Satzgrenze	X_1	...	X_n	finite Verbform	Y_1	...	Y_n	Satzgrenze

In Bezug auf das Deutsche können die Wortformen Y_1 – Y_n Mittelfeldeinheiten, rechten Satzklammerteilen, Nachfeldeinheiten oder rechten Außenfeldeinheiten zugeordnet werden. Ihre Analyse steht nicht im Mittelpunkt dieser Untersuchung. Dem Textbereich X_1 – X_n lassen sich Einheiten des linken Außenfelds oder Vorfeldeinheiten zuordnen, sofern kein Verbletztsatz vorliegt (vgl. (5)).

Für die quantitative kontrastive Korpusuntersuchung bietet sich ein abstrakteres Modell an, das linke Felder als sog. N-Gramme beschreibt. Bei einem N-Gramm handelt es sich um eine beliebig lange Folge von sprachlichen Einheiten, wobei N für die jeweilige Anzahl solcher Einheiten steht. In Bezug auf linke Felder können je nach Fragestellung lexikalische Einheiten, z.B. Wortformen (Tokens), Lexeme (Types) oder wortartenklassifizierte Parts-of-Speech (POS) Ausgangspunkt der Betrachtung sein. Aus kombinatorischen Gründen erscheint für den Sprachvergleich die Wortarten-/POS-Ebene am sinnvollsten, da die Anzahl der zur Verfügung stehenden POS-Tags in den Tagsets der untersuchten Korpora begrenzt und die Zahl der Realisierungsmöglichkeiten eines N-Gramms somit überschaubar ist. In dieser Korpusuntersuchung ist die konstituierende sprachliche Einheit, sofern nicht anders angegeben, zunächst einfach das „Wort", wobei sich die Definition des Wortbegriffs bei der Korpusanalyse aus der default-Tokenisierung der Kontrastkorpora (s. Kap. 3.3) ergibt. Ein linkes Feld, das aus zwei Wörtern besteht, kann also als Bigramm verstanden werden, eines mit drei Wörtern als Trigramm usw. Bei einer Segmentierung nach N-Grammen werden wiederum Types (im Sinne der üblichen Type-Token-Unterscheidung) generiert, die auf der Größe der jeweiligen linken Felder basieren, wobei N für die Anzahl der Wörter des entsprechenden linken Felds steht. Die sich auf die linken Felder beziehenden Types und Tokens sind nicht mit Wortform-Types/-Tokens zu verwechseln. In

dieser Untersuchung wurden alle Vorkommen (Tokens) eines Types von linken Feldern (z.B. ein linkes Feld mit genau einem oder genau fünf Wörtern usw.) in den Korpora gezählt und kontrastiv ausgewertet.

Die Auswertung der Korpusdaten soll Aufschluss darüber geben, ob es einen Zusammenhang zwischen Größe/Häufigkeitsverteilung von linken Feldern, den Struktureigenschaften der jeweiligen Sprache und der Textsorte gibt (s. Kap. 4).

3.3 Korpusspezifische Suchanfragen zur Ermittlung der linken Felder

Die Analyse von linken Feldern unter morphosyntaktischen oder funktional-semantischen Gesichtspunkten erfordert spezifische Annotationsebenen. In den Wikipedia-Korpora sind detailliertere morphosyntaktische Informationen zu nominalen Elementen für einfache grammatische Kategorisierungen wie Genus, Numerus und Kasus zwar teilweise verfügbar, allerdings nicht in allen sechs Kontrastkorpora und nicht in vergleichbarer Annotationstiefe. Die nach Parts-of-Speech annotierten Daten erlauben deshalb nur auf der Wortarten-Ebene eine sprachübergreifende Analyse. Die gegebenen POS-Tags können genutzt werden, um die Begrenzung der linken Felder durch das finite Verb oder einen äquivalenten Ausdruck in den Korpus-Suchanfragen zu definieren und um die sich darin befindenden Wörter hinsichtlich ihrer Wortartenzugehörigkeit auszuwerten.

In dieser Untersuchung unberücksichtigt bleiben die hierarchischen Strukturen von Ober- und Untersätzen. Die Suchanfragen liefern nicht nur linke Felder von Hauptsätzen, sondern auch solche von Nebensätzen, wenn der als Satz annotierte Textbereich mit einem subordinierten Teilsatz beginnt. Die Sichtung der Suchergebnisse hat aber gezeigt, dass vorangestellte Nebensätze in den untersuchten Korpora nur selten vorkommen.

Die Satz- und Wortgrenzen in den Korpora, auf denen die Suchanfragen operieren, sind in COSMAS II durch eine für alle Kontrastkorpora identische Tokenisierung definiert, die auch dem Deutschen Referenzkorpus (DeReKo) zugrundeliegt. Die komplex strukturierten Suchanfragen zur Ermittlung der linken Felder in den Wikipedia-Korpora sind so aufgebaut, dass alle auswertbaren[7] Vorkom-

7 Auswertungsprobleme gibt es mit linken Feldern, die bestimmte unerwünschte Suchausdrücke enthalten. Es handelt sich dabei um Überbleibsel von Formatierungsausdrücken aus den Wikipedia-Rohdaten (v.a. bestimmte Interpunktionszeichen, z.B. „{{“ zur Markierung von Ankern im Text), die beim Konvertieren in das DeReKo-Format nicht korrekt entfernt wurden und in Folge unsinnige POS-Annotationen nach sich ziehen. Diese Sätze oder linken Felder sind in

men erfasst werden, die maximal 50 Wörter[8] umfassen. Als Begrenzung für das linke Feld dienen die POS-Tags, mit denen finite Verben bzw. äquivalente Ausdrücke[9] in den jeweiligen Sprachen erfasst werden. Das zu ermittelnde linke Feld beginnt am Satzanfang (erstes Wort eines Satzes) und endet unmittelbar vor dem ersten begrenzenden Ausdruck. Das letzte Wort eines linken Feldes ist also das Wort unmittelbar vor dem finiten Verb. Die Suchanfragen wurden jeweils korpusspezifisch modifiziert, da in allen sechs Sprachen unterschiedliche Tagsets vorliegen. Als Resultat erscheinen Tokens von linken Feldern, die sich aus einer jeweils unterschiedlichen Anzahl von Wörtern (Worttokens) zusammensetzen.

Das ungarische Wikipedia-Korpus stellt sowohl den Umfang als auch das Recherche- und Analysesystem betreffend einen Sonderfall dar. Im Gegensatz zu den anderen Korpora beinhaltet es nur eine ca. 10 Millionen Wörter (MW) umfassende Zufallsauswahl an Texten, nicht die gesamte zum Zeitpunkt der Korpuserstellung verfügbare Wikipedia. Der entscheidende Unterschied für die Entwicklung und Durchführung der Suchanfragen ist aber die Tatsache, dass die ungarischen morphosyntaktischen Annotationen in COSMAS II nicht vollständig implementiert sind. Die Recherchen im ungarischen Wikipedia-Korpus wurden deshalb mit Hilfe von Anwendungen der IMS Open Corpus Workbench (CWB) durchgeführt, die Abfrage erfolgte über die web-basierte grafische Benutzeroberfläche CQPweb. Die hier verwendete Abfragesprache „CQP" funktioniert im Unterschied zu COSMAS II auf der Basis von regulären Ausdrücken, was eine abweichende Konzeption der Suchanfragen erforderte.

3.4 Worttokens und Types von linken Feldern

Wie oben beschrieben, erfassen die Suchanfragen linke Felder von unterschiedlicher Größe (d.h. Types von linken Feldern) und geben Auskunft über deren jeweilige absolute Häufigkeit im Korpus (Tokenfrequenz von linken Feldern). Die Grundlage der Gesamttrefferanzahl bildet also die Anzahl der Sätze, die linke Felder nach den in der Suchanfrage spezifizierten Kriterien aufweisen, und nicht die

---

den meisten Fällen unbrauchbar. Da sie in statistisch relevantem Umfang vorliegen, verbessert sich die Qualität der Suchergebnisse erheblich, wenn sie von vorneherein ausgeschlossen werden.

8 Die maximale Grenze ist willkürlich gesetzt, basiert aber auf Erfahrungen im Umgang mit den Korpusdaten. Mehr als 50 Tokens umfassenden Suchergebnissen liegen zu häufig fehlerhafte Annotationen zugrunde, um sie in die Analyse miteinbeziehen zu können.

9 Für das Polnische wurden unpersönliche Bildungen auf *-no/-to* (s. Przepiórkowski/Woliński 2003) als Äquivalente zu finiten Verben herangezogen.

Anzahl der Wörter, die innerhalb dieser linken Felder anzutreffen sind. Die Gesamtanzahl aller Worttokens innerhalb von linken Feldern einer bestimmten Größe entspricht dem Produkt aus ihrer absoluten Häufigkeit und ihrer Größe gemessen in Anzahl der Wörter. So zählen z.B. 100 linke Felder à zwei Wörter in der Summe 200 Worttokens, die gleiche Anzahl von linken Feldern mit je drei Wörtern folglich 300 Worttokens usw. Am konkreten Beispiel wird ein besonderer Effekt deutlich: Die Suchanfrage im deutschen Wikipedia-Korpus lieferte genau 4.134.543 linke Felder à zwei Wörter; die Anzahl der darin vorkommenden Worttokens beträgt also 8.269.086. Im Korpus befinden sich demnach ca. 8,27 MW, die diesen Type von linkem Feld zugeordnet werden können. Zum Vergleich: Im selben Korpus beträgt die Anzahl der linken Felder à drei Wörter nur 2.929.899 und ist damit deutlich geringer als jene à zwei Wörter. Dennoch befinden sich im Korpus mehr Wörter, die dem linken Feld mit drei Wörtern zuzuordnen sind (8,79 MW), als Wörter, die im linken Feld mit zwei Wörtern anzutreffen sind (8,27 MW). Würde man ein beliebiges Wort aus einem linken Feld nehmen, so wäre also die Wahrscheinlichkeit, dass dieses Worttoken zu einem linken Feld mit drei Wörtern gehört, höher als dessen Auftretenswahrscheinlichkeit in einem mit zwei Wörtern, obwohl linke Felder mit drei Wörtern seltener sind als solche mit zwei Wörtern. Dieser statistische Effekt macht sich bei der Darstellung der Häufigkeiten in den Korpora bemerkbar (vgl. Abbildung 1 mit Abbildung 2 in Kap. 4.1).

3.5 Vereinheitlichungen für alle sechs Kontrastkorpora

Die Wortartenklassifizierung geht auf die einzelsprachlichen POS-Tags der morphosyntaktisch annotierten Korpora zurück, die aber im Rahmen dieser Untersuchung zum Zwecke der Vergleichbarkeit teilweise zu übereinzelsprachlichen Wortklassen zusammengefasst wurden. Da zum Ermitteln von Phrasenstrukturen eine gröbere, wortartenbasierte Klassifizierung ausreicht, wurden im Post-Processing differenzierte Subklassifizierungen von POS-Tags nicht berücksichtigt, bzw. nur einzelne Merkmale als Grundlage für eine vereinfachte und übereinzelsprachliche Neuklassifizierung herangezogen (vgl. Tabelle 2):

Tab. 2: Beispiele für Neuklassifizierungen auf der Basis von POS-Tags in den Korpora

Korpus	POS-Tags (Beispiele)	übereinzelsprachliche Neuklassifizierung
DE	für <AP pr> uns <PRON per irr>	für <P> uns <PERS>
FR	la <DET art> population <NOM> du <PRP pdet> Mexique <NAM>	la <Art> population <N> du <P> Mexique <N>
IT	Un <DET indef> ' <PON> arma <NOM> inastata <VER pper>	Un' <Art> arma <N> inastata <Part>
NO	En <DET mask ent kvant> annen <DET dem mask ub ent @adj> anarkistisk <ADJ ub m/f ent pos> forfatter <SUBST appell mask ub ent>	En <Q> annen <Dem> anarkistisk <Adj> forfatter <N>
PO	2 <TNUM integer> tygodnie <SUBST pl acc m3>	2 <Num> tygodnie <N>
UN	A <Det> lagúnák <N.PL.NOM> tengelyében <N.PSe3.INE>	A <Art> lagúnák <N> tengelyében <N>

Welche Schwierigkeiten sich aus den einzelsprachlich unterschiedlichen Klassifizierungsmodellen für den Vergleich von formal gleichen sprachlichen Ausdrücken ergeben, lässt sich sehr gut am Beispiel der Numeralia illustrieren (vgl. Tabelle 3[10]).

10 Die POS-Tag Zuweisung von als Ziffer(n)+Punkt dargestellten Ordinalzahlen wurde im frz. Wikipedia-Korpus völlig uneinheitlich vorgenommen. Der Tagger wurde offenkundig nicht auf die Erkennung dieses Numeraletyps trainiert.

Tab. 3: POS-Tags für Numeralia in den Wikipedia-Korpora vor der Vereinheitlichung

Numerale		**korpusspezifische POS-Tag-Zuweisung**					
Form	**Bsp.**	**DE**	**FR**	**IT**	**NO**	**PO**	**UN**
Ziffer(nfolge)	*41*	CARD	NUM	NUM	DET kvant	TNUM integer	DIG
Ziffer(nfolge) + Punkt	*7.*	ADJ at	*fehlerhaft*[10]		ADJ @ ordenstall	TNUM integer + INTERP	NUM.Nom
Ziffernfolge mit Komma	*21,4*	CARD	NUM		DET kvant	TNUM frac	DIG
Bruchzahl	*2/3*					TSYM	DIG
Datum	*2.6.1980*						–
Prozentangabe	*3%*	CARD + N nn	NUM + SYM	NUM+SYM			DIG
Kardinalzahl	*zwei*	CARD	NUM	ADJ		NUM kasus	NUM.Kasus
	eins		NUM / DET art	ADJ / NOM		ADJ kasus	
Ordinalzahl	*dritte*	ADJ at	ADJ / NOM	ADJ / DET indef / NOM	ADJ @ ordenstall	ADJ kasus	
Bruchzahl	*drittel*	ADJ at / N nn			SUBST appell		

Im Rahmen der Neuklassifizierung wurden die verschiedene Numeralia markierenden Tags zu einer gemeinsamen Klasse Num (für Numerale) zusammengefasst. In der norwegischen Sprachversion umfassen POS-Tags, die DET kvant enthalten, allerdings nicht nur Numeralia, sondern auch den Indefinitartikel. Ihnen wurde der Kürze halber als neue Bezeichnung Q (für Quantor) zugewiesen. Auf eine Zuweisung des Tags Art (für Artikel) zum Indefinitartikel wurde für das Norwegische verzichtet, da es aufgrund des Fehlens des korrespondierenden Definitartikels diesbezüglich gesondert betrachtet werden sollte (s. 4.2.1). Insbesondere bei der Klassifizierung der Kardinalzahlen bleibt aber auch nach der übereinzelsprachlich vorgenommenen Vereinheitlichung eine Diskrepanz zwischen den Korpora bestehen, in denen Zahlwörter entweder als Numeralia oder als Adjektive, Nomina oder Artikel gekennzeichnet sind. Die Ziffer(nfolge) – der frequenteste Numeraletyp in den Korpora – wurde hingegen durchgängig zu Num vereinheitlicht. Datumsangaben vom Typ TT.MM.JJJJ sind im Korpus vergleichsweise selten. Das deutschsprachige Wikipedia-Autorenportal empfiehlt, den Monat als Wort auszuschreiben[11] und behält sich vor, Artikel mit abweichenden Datumsformaten (Zitate ausgenommen) als korrekturbedürftig aufzulisten (Wikipedia 2014b). Die anderen fünf kontrastsprachlichen Wikipedias geben diesbezüglich vergleichbare Richtlinien vor.[12]

Auch auf der Ebene der Worttokens wurden weitreichende vereinheitlichende Änderungen vorgenommen. Bei Ziffern(folgen) wurde ihrer allgemeinen semantischen Eigenschaft, einen Ausdruck zu quantifizieren, mehr Gewicht beigemessen als ihrer Eigenschaft, einen spezifischen Ziffern-/Zahlenwert auszudrücken. In allen sechs Kontrastkorpora wurden deshalb Ziffern und Ziffernfolgen durch einen Platzhalter ersetzt, sodass sämtliche Zahlen und verschiedene Ziffern/Zahlen enthaltenden Ausdrücke zu jeweils einem Type zusammengefasst werden konnten. Um auch die in allen Korpora frequenten Datumsangaben auf dieser Analyseebene miteinbeziehen zu können, wurde mit Monats- und Tagesnamen ebenso verfahren. Der gewünschte Effekt der Maßnahme zeigt sich in den Ranglisten von Wortformen-N-Grammen: N-Gramme, die Zahlen oder Monats-/Tagesnamen als Bestandteil haben, tauchen in der Frequenz-Rangliste von Wortformen weiter oben auf, da sie unabhängig von ihrem Wert zu einem N-Gramm zusammengefasst werden, z.B. im Korpus: *Im Jahr 1990*; *Im Jahr 2014*, im Post-

11 Dadurch erscheinen Datumsangaben in der übereinzelsprachlichen Wortklassifizierung i.d.R. als Abfolge von Num+N+Num.

12 Das Datumsformat „TT. Monatsname JJJJ" wird im Ungarischen durch „JJJJ. Monatsname TT." notiert, da die Monats-/Jahresangaben ursprünglich Attribute der Tagesangaben waren und als solche im Ungarischen typischerweise vorangestellt werden (Pilarský 2013: 187f.).

Processing: *Im Jahr* {Zahl}; *Am 1. Januar*; *Am 28. Juni* → *Am* {Zahl}. {Monat}. Denselben Zweck verfolgen einige sprachspezifische Vereinheitlichungen auf der Wortformen-Ebene: Bestimmte morphophonologische Varianten von Funktionswörtern/Klitika (z.B.: frz. *se*; *s'* → *s(e/')*; it. *nel*; *nell'*→ *nel(l')*) oder Affixen (z.B. ung. *-ban/-ben*) wurden ebenfalls zusammengefasst, um das Ausufern von N-Grammen zu vermeiden.

Kein Eingriff wurde in den folgenden Fällen vorgenommen: Mit Bindestrich und ohne Spatium angegebene Zeitspannen {Zahl}-{Zahl} (z.B. *2001–2011*) werden in den Korpora wie ein einziges Token behandelt und liegen entsprechend annotiert vor, obwohl dieser Schreibkonvention eigentlich ein dreigliedriger Ausdruck zugrunde liegt (*2001 bis 2011*). Aufgrund der weitreichenden Konsequenzen, die eine erneute Tokenisierung und Klassifizierung im Post-Processing nach sich zöge, wurde angesichts der geringen Frequenz dieser Ausdrücke auf eine solche Korrekturmaßnahme verzichtet.[13] In den deutschen Wikipedia-Korpusdaten beträgt der Anteil von Zeitspannenangaben im Verhältnis zu Angaben mit einfacher Jahreszahl nur 0,5%.

4 Ergebnisse

4.1 Types von linken Feldern

Für einen aussagekräftigen Überblick genügt bereits ein Vergleich der relativen Häufigkeiten von linken Feldern mit der Größe von einem bis zwölf Wörtern. Sie sind für alle sechs Wikipedia-Korpora in Abbildung 1 dargestellt. Während sich die Häufigkeitsverteilung der größeren linken Felder (≥7) in den sechs kontrastsprachlichen Korpora grundsätzlich ähnelt, treten im Bereich der kleineren deutliche Unterschiede hervor. Anhand der relativen Häufigkeit der Tokens, die einem Type von linkem Feld zugeordnet werden, d.h. ihrer Ranghäufigkeitsverteilung, können die sechs Kontrastkorpora in zwei Gruppen unterteilt werden: Es gibt Korpora, bei denen Größe und Ranghäufigkeitsverteilung (Types) von linken Feldern von Beginn an übereinstimmen (DE/FR/NO); und solche, bei denen diese negative Korrelation (je größer desto seltener) erst ab einer bestimmten Größe eintritt (IT/PO/UN). Für das deutsch-, französisch- und norwegischsprachige Wikipedia-Korpus gilt der Grundsatz: Je größer das linke Feld, desto geringer die Häufigkeit der entsprechenden Tokens. Für das italienisch-, polnisch- und unga-

13 Das betrifft auch Ausdrücke mit anderen Interpunktionszeichen, z.B. 12/1996.

rischsprachige Korpus gilt dieser Grundsatz erst ab den Rängen 4 bzw. 7, wie Tabelle 4 entnommen werden kann:

Tab. 4: Ranghäufigkeitsverteilung der Types von linken Feldern (W=Wort im linken Feld)

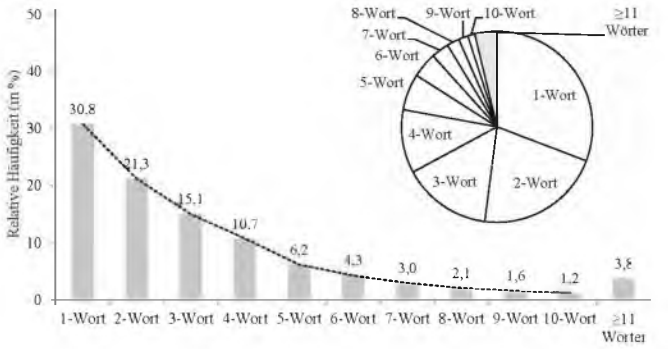
Häufigkeitsrang	Korpusgruppen		
	DE / FR / NO	IT / PO	UN
1.	W_1	W_1_W_2	W_1_W_2_W_3
2.	W_1_W_2	W_1_W_2_W_3	W_1_W_2_W_3_W_4
3.	W_1_W_2_W_3	W_1	W_1_W_2_W_3_W_4_W_5
4.	W_1_W_2_W_3_W_4	W_1_W_2_W_3_W_4	W_1_W_2
5.	W_1_W_2_W_3_W_4_W_5	W_1_W_2_W_3_W_4_W_5	W_1_W_2_W_3_W_4_W_5_W_6
6.	W_1_W_2_W_3_W_4_W_5_W_6	W_1_W_2_W_3_W_4_W_5_W_6	W_1

In der ersten Gruppe sind Sätze mit linken Feldern, die genau ein Wort umfassen, am häufigsten. Statistisch betrachtet besitzt in Wiki-NO sogar jeder zweite Satz (50,1%) ein linkes Feld, das aus nur einem Wort besteht (vgl. Tabelle 5). In Wiki-DE ist das knapp jeder dritte Satz (30,8%), in Wiki-FR noch knapp jeder vierte Satz (23,54%). Ein Blick auf die relative Häufigkeit der Tokens von linken Feldern (Abbildung 1) lässt in diesen drei Korpora einen charakteristischen hyperbelartigen Verlauf der Verteilungskurve erkennen. Die zweite Gruppe umfasst das polnisch- und italienischsprachige Korpus: Hier sind linke Felder mit zwei Wörtern am häufigsten. Diese Präferenz ist in Wiki-IT im Verhältnis allerdings deutlich stärker ausgeprägt als in Wiki-PO, wo linke Felder mit einem bis drei Wörtern ähnlich häufig vorkommen (vgl. Abbildung 1/Tabelle 5). Wiki-UN stellt einen Einzelfall dar. Sehr kleine Felder mit einem oder zwei Wörtern sind hier seltener, das linke Feld mit genau drei Wörtern ist am frequentesten, wobei linke Felder mit drei bis fünf Wörtern ähnlich häufig vorkommen.

Die folgenden Abbildungen 1 und 2 zeigen korpusspezifische Verteilungskurven für ein bis zwölf Wörter große linke Felder. Die Werte geben relative Häufigkeiten an, die sich in Abbildung 1 (a bis f) auf die Anzahl aller Tokens von linken Feldern und in Abbildung 2 (a bis f) auf die Anzahl aller Wörter (Worttokens) beziehen.

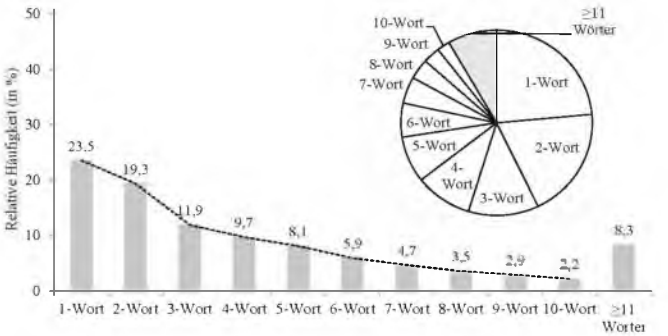
Abb. 1a–f: Relative Häufigkeit (in %) von Linke-Feld-Types in den Wikipedia-Korpora

a Wiki-DE



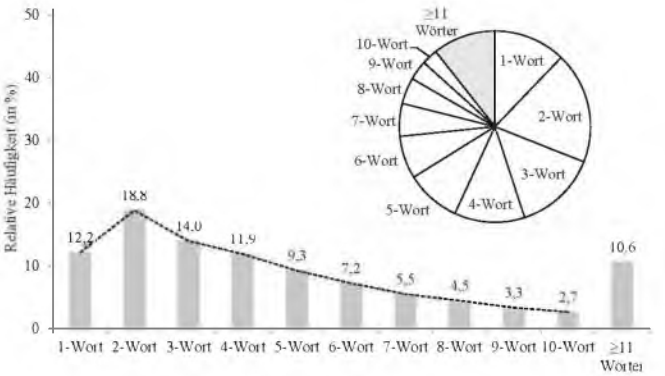
Linkes-Feld-Type (nach Anzahl der Worter)

b Wiki-FR



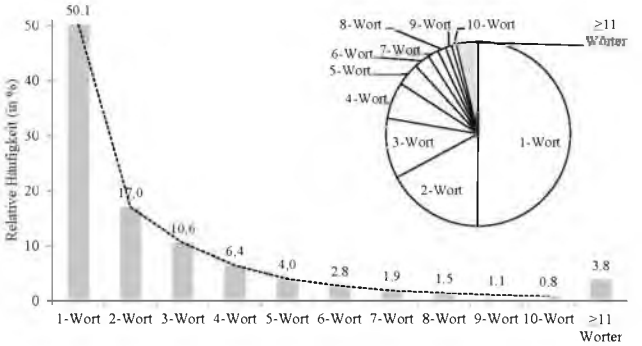
Linkes-Feld-Type (nach Anzahl der Worter)

c Wiki-IT



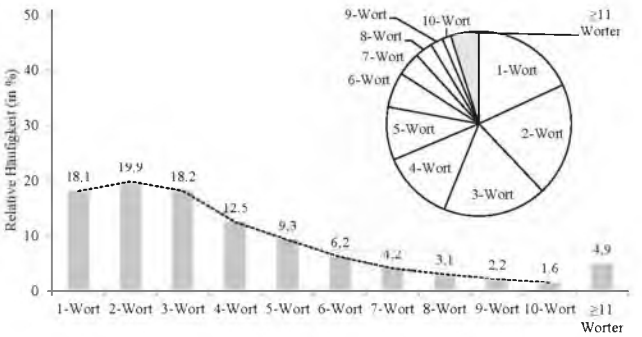
Linkes-Feld-Type (nach Anzahl der Worter)

d Wiki-NO



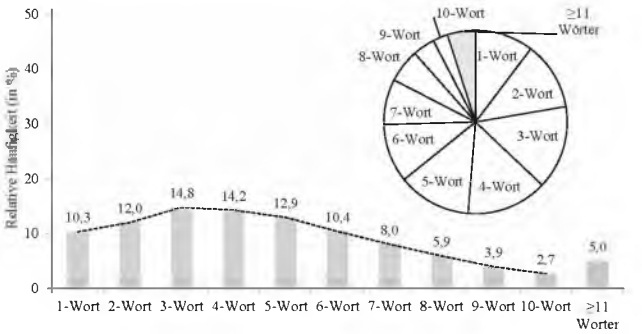
Linkes-Feld-Type (nach Anzahl der Worter)

e Wiki-PO



Linkes-Feld-Type (nach Anzahl der Worter)

f Wiki-UN



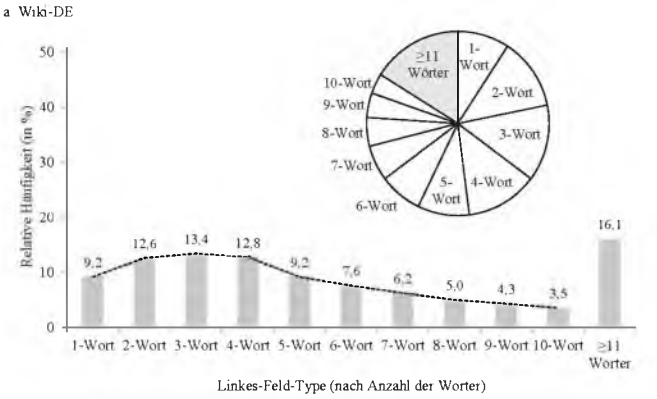
Linkes-Feld-Type (nach Anzahl der Worter)

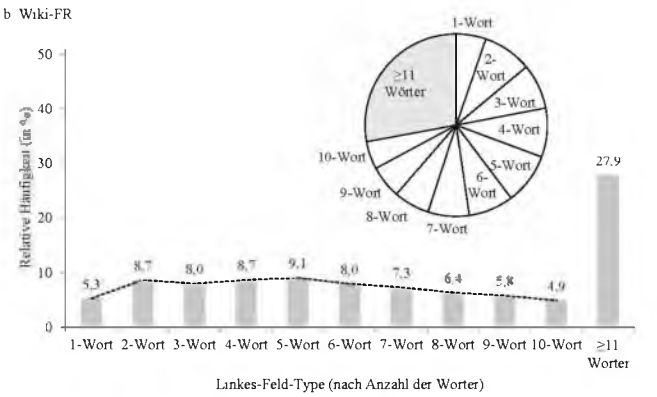
Abbildung 1 gibt die Daten in Tabelle 5 wieder:

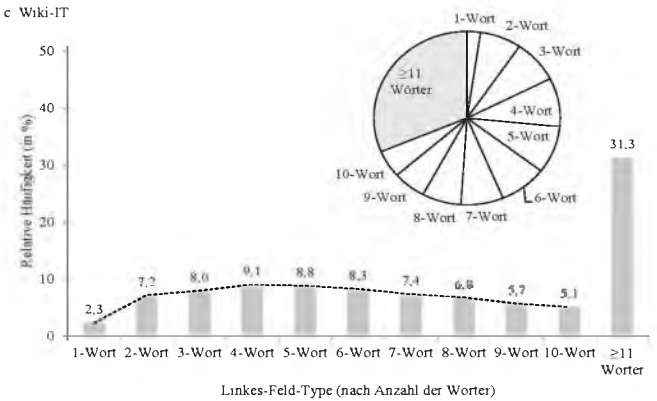
Tab. 5: Relative Häufigkeit (in %) von linken Feldern in den Wikipedia-Korpora

Größe linkes Feld	Korpus					
	DE	FR	IT	NO	PO	UN
1	30,8	23,5	12,2	50,1	18,1	10,3
2	21,3	19,3	18,8	17	19,9	12
3	15,1	11,9	14	10,6	18,2	14,8
4	10,7	9,7	11,9	6,4	12,5	14,2
5	6,2	8,1	9,3	4	9,3	12,9
6	4,3	5,9	7,2	2,8	6,2	10,4
7	3	4,7	5,5	1,9	4,2	8
8	2,1	3,5	4,5	1,5	3,1	5,9
9	1,6	2,9	3,3	1,1	2,2	3,9
10	1,2	2,2	2,7	0,8	1,6	2,7
11	0,9	1,7	2	0,7	1,1	1,7
12	0,7	1,3	1,6	0,5	0,9	1,2
13	0,5	1,1	1,3	0,4	0,6	0,7
14	0,4	0,9	1	0,4	0,5	0,5
15	0,3	0,7	0,8	0,3	0,4	0,3
16	0,2	0,5	0,7	0,2	0,3	0,2
17	0,2	0,4	0,6	0,3	0,2	0,1
18	0,1	0,3	0,4	0,2	0,2	0,1
19	0,1	0,3	0,4	0,1	0,1	0,1
20	0,1	0,2	0,3	0,1	0,1	0
Summe 21–50	0,3	0,9	1,4	0,6	0,4	0,1

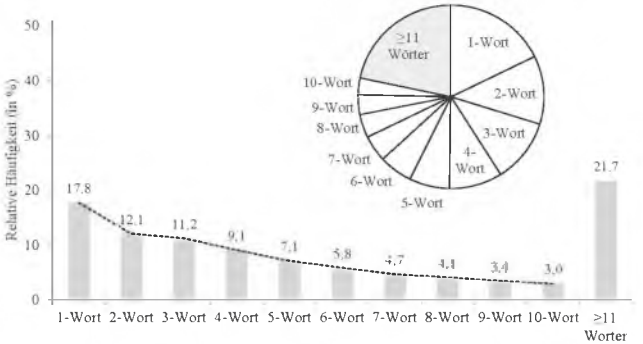
Abb. 2: Relative Häufigkeit von Worttokens (in %) in den Wikipedia-Korpora





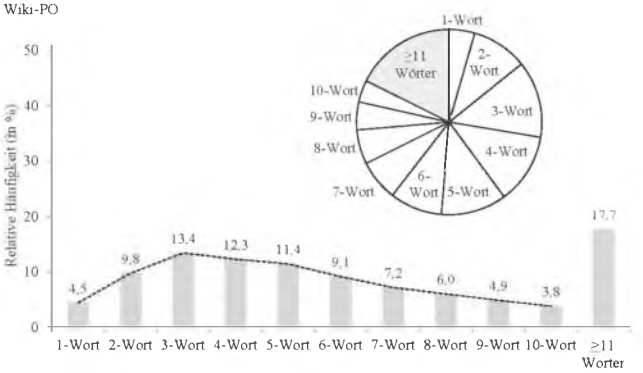


d Wiki-NO



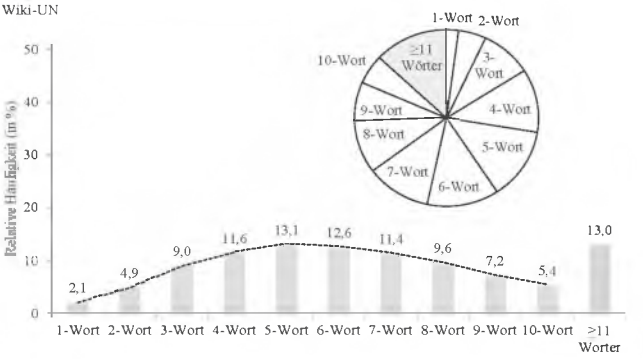
Linkes-Feld-Type (nach Anzahl der Wörter)

e Wiki-PO



Linkes-Feld-Type (nach Anzahl der Wörter)

f Wiki-UN



Linkes-Feld-Type (nach Anzahl der Wörter)

4.2 Korrelationen zwischen Phrasenstruktur und Größe der linken Felder

Die Vermutung, dass es einen Zusammenhang zwischen Größe/Häufigkeitsverteilung von linken Feldern und den Struktureigenschaften der jeweiligen Sprache gibt, wird durch gezielte Überlegungen wie etwa zu Realisierungsformen (Phrasenstruktur) und Stellung von Komplementen/Supplementen oder zu Konnektoren in den Kontrastsprachen gestützt. Betrachtet man Phrasen als N-Gramme von Wörtern, die bestimmten Wortklassen zugeordnet werden können, lassen sich eindeutige Zusammenhänge zwischen Größe und Wortklassenbesetzung von linken Feldern einerseits und bestimmten Phrasentypen andererseits herstellen. Zur Ermittlung der Verteilung von Wortklassen wurden in allen Korpora gleich große Stichproben von Vorkommen linker Felder desselben Umfangs genommen.

In linken Feldern von einem bis vier Wörtern sind die häufigsten Phrasentypen Nominal-, Pronominal- und Präpositionalphrasen (s. Tabelle 6[14] und Tabelle 7a/b). Bei größeren linken Feldern nimmt die Varianz der N-Gramme in Bezug auf die Linearisierung ihrer Elemente so stark zu, dass ohne ein syntaktisches Parsing keine zuverlässigen Rückschlüsse auf die Phrasenstrukturen dieser linken Felder getroffen werden können. Die folgenden Tabellen geben die prozentuale Verteilung von Wortklassen innerhalb der Types von linken Feldern für alle sechs kontrastsprachlichen Wikipedia-Korpora geordnet nach Häufigkeitsrängen wieder. Da die Datengrundlage auf Korpusstichproben basiert, sind Ränge, die aufgrund von statistischen Berechnungen[15] nicht eindeutig zugeordnet werden konnten, teilweise mehrfach besetzt. Bei zu kleinen Werten kann aus denselben Gründen i.d.R. gar keine sinnvolle Rangzuweisung mehr vorgenommen werden. Angegeben wurden die ersten, zweiten und dritten Ränge, sofern ihre relative Häufigkeiten 0,5% nicht unterschreiten.

14 Abkürzungen: Adj – Adjektiv; Adv – Adverb; Art – Artikel; Dem – Demonstrativum; N – Nomen; Num – Numerale; P – Adposition; Pers – Personalpronomen; Pro – sonstiges Pronomen; Q – Quantor

15 Hierzu wurden für alle Häufigkeitsränge Konfidenzintervalle berechnet. Ränge mit sich überlappenden Intervallen wurden zusammengefasst.

Tab. 6: Relative Häufigkeit (in %) von Wortklassen-N-Grammen

Linkes Feld mit	Rang	Wikipedia-Korpus DE		FR		IT		NO		PO		UN	
einem Wort	1.	Pers	25,5	Pers	71,8	N	46,0	N	56,4	N	65,4	Adv	26,7
	2.	N	20,4	N	23,6	Pers	17,0	Pers	36,1	Adv	19,2	Num	21,0
	3.	Adv	16,7	Pro	2,9	Adv	13,9	Adv	2,8	Adj	5,2	N	18,3
zwei Wörtern	1.	Art+N	45,4	Art+N	36,6	Art+N	46,4	N+N	28,4	N+N	18,1	Num+N Art+N	12,1 11,7
	2.	P+N	11,2	Pro+N N+N	13,6 13,5	P+Num	10,2	P+N	16,8	N+Adj	14,5	Adj+N	7,4
	3.	Pro+N	10,5	Pers+Pers	11,2	Pro+N	7,9	P+Q	14,9	P+N	14,1	N+N	5,1
drei Wörtern	1.	Art+Adj+N	16,3	Art+N+Adj	13,8	Art+N+Adj	11,5	N+P+N Dem+Adj+N	13,7 13,5	P+Num+N	13,9	Art+N+N Art+N+Adv	6,0 5,6
	2.	P+Art+N	13,7	P+Num+Pers	10,8	Art+N+N	7,3	N+N+N	8,6	P+N+N	8,9	Pro+Art+N	3,8
	3.	P+N+Num	12,1	Art+N+N	7,9	Art+Adj+N	6,1	P+Dem+N	4,6	P+N+Num	8,1	Art+Adj+N Num+Adj+N Num+N+Num Num+Num+N	3,1 3,0 3,0 2,8
vier Wörtern	1.	Art+N+Art+N	11,1	Art+N+P+N	23,7	Art+N+P+N	20,4	P+N+P+N	8,9	P+N+Num+N	7,4	Art+N+Adj+N	3,8
	2.	P+Art+Adj+N	9,7	P+Num+Art+N	3,7	Art+Num+N+Num	2,9	P+Dem+Adj+N	4,1	P+Num+N+N	5,0	Art+N+Num+N	3,1
	3.	P+Adj+N+Num	9,5	Art+N+N+N	2,5	P+N+P+N	2,8	Dem+Adj+N+Q	3,2	P+N+N+N	4,6	Art+Adj+N+N	2,3

Tab. 7a: Relative Häufigkeit (in %) von Wortformen-N-Grammen (Deutsch, Französisch, Italienisch)

Linkes Feld mit	Rg.	Wikipedia-Korpus					
		DE		FR		IT	
einem Wort	1.	#	13,4	*Il*	45,7	*Si*	16,2
	2.	*Er*	10,9	*Elle*	14,1	*Non*	3,2
	3.	*Sie*	7,5	*On*	4,9	*Questo*	2,2
zwei Wörtern	1.	*Seit #*	2,5	*Il s(e/')*	3,9	*Nel #*	9,2
	2.	*Ab #*	1,5	*Il y*	1,9	*Il film*	0,9
	3.	*Bis #*	0,8	*Elle s(e/')*	1,3	*Dal #*	0,7
		Darüber hinaus	0,7	*Il n(e/')*	1,2	*L'album*	0,7
		Der Ort	0,6			*La città*	0,6
drei Wörtern	1.	*Im Jahr(e) #*	5,4	*En # (') il*	8,0	*Scoperto nel(l') #*	1,9
	2.	*Im* {Monat} #	3,9	*En # (') elle*	1,5	*Nel(l') # si*	1,3
	3.	*Am #.* {Monat}	0,6	*		*Nel(l')* {Monat} #	0,9
		Im #. Jahrhundert	0,6				
		# bis #	0,6				
		In dieser Zeit	0,5				
		Im selben Jahr	0,5				
vier Wörtern	1.	*Am #.* {Monat} #	8,9	*En* {Monat} #') *il*	1,1	*Il/l' #* {Monat} #	3,0
	2.	*Von # bis #*	5,0	*En # (') il s(e/')*	0,5	*Dal(l') # al(l') #*	1,2
	3.	*Zwischen # und #*	1,6	*		*Nel(l')* {Monat} *del(l') #*	0,6

steht für eine beliebige Zahl

* Aufgrund zu kleiner Werte (ca. <0,5%) kann keine eindeutige Rangzuweisung vorgenommen werden.

Tab. 7b: Relative Häufigkeit (in %) von Wortformen-N-Grammen (Norwegisch, Polnisch, Ungarisch)

Linkes Feld mit	Rg.	Wikipedia-Korpus					
		NO		PO		UN	
einem Wort	1.	*Han*	13,5	*Nie*	3,3	*#-b(a/e)n*	18,3
	2.	*Det*	8,5	*Następnie*	1,7	*Itt*	4,4
				Obecnie	1,7		
	3.	*Den*	4,4	*Początkowo*	0,9	*Ez*	3,3
				Miasto	0,8		
				Później	0,8		
				Miejscowość	0,7		
zwei Wörtern	1.	*I #*	13,7	*W #*	9,0	# {Monat} *(j)(á/é)ban*	1,2
	2.	*I tillegg*	1,5	*Od #*	1,3	{Monat} *#-(j)(á/é)n*	0,5
						Cornelis	0,4
						Johannes	
	3.	*I Norge*	1,1	# {Monat}	0,5	*	
		I dag	0,9	*Rok później*	0,4		
drei Wörtern	1.	*Den #.* {Monat}	0,6	*W # roku*	9,7	#. {Monat} *#-(j)(á/é)n*	2,4
	2.	*Personene i listen*	0,5	*W latach #-#*	2,7	*#-b(a/e)n* *# lakosa*	1,6
	3.	*		# {Monat} #	2,3	*	
vier Wörtern	1.	*Den #.* {Monat} #	3,6	# {Monat} # *roku*	2,1	*	
	2.	*Fra # til #*	2,2	*W latach #-#* *miejscowość*	1,7	*	
	3.	*De andre på laget*	0,8 0,8	*W* {Monat} # *roku*	1,6	*	
		Mellom # og					

steht für eine beliebige Zahl

* Aufgrund zu kleiner Werte (ca. <0,5%) kann keine eindeutige Rangzuweisung vorgenommen werden.

Die Verteilung der N-Gramme in den Kontrastsprachen ist insbesondere für die Wortklassen Artikel, Präposition und Personalpronomen aussagekräftig, aus denen sich verschiedene Ausprägungen bestimmter Phrasentypen ableiten lassen, die auch in Verbindung mit bestimmten syntaktischen Funktionen gebracht werden können. Die Auswahl von aus kontrastiver Sicht relevanten Phänomenen, auf die im folgenden Abschnitt näher eingegangen wird, ergibt sich unmittelbar aus den Daten in Tabelle 6 und Tabelle 7a/b. Das Spektrum der Erkenntnisse erstreckt sich dabei über die unterschiedlichen Ebenen der Wortklassen

(Pronomen, Kap. 4.2.4), Phrasentypen (Nominalphrase, Kap. 4.2.1), funktionale Ausdrucksklassen (Temporaladverbiale, Kap. 4.2.2) und Besetzungsrestriktionen (mehrere Phrasen im linken Feld, Kap. 4.2.3) unter Berücksichtigung ihrer Textsortenspezifizität (Kap. 4.2.5).

4.2.1 Nominalphrasen

In der deutschen, französischen, italienischen und ungarischen Wikipedia sind linke Felder mit zwei Wörtern am häufigsten durch Nominalphrasen (NPs) mit Artikel besetzt, wobei es sich um die typische Realisierungsform und Stellung des Subjekts handelt. Im Deutschen und Ungarischen können an dieser Stelle und in dieser formalen Ausprägung auch andere Komplemente, die nicht das Subjekt repräsentieren, erscheinen. Nach Sichtung der Belege zeigt sich, dass im deutschsprachigen Wikipedia-Korpus von dieser Möglichkeit jedoch kaum Gebrauch gemacht wird. Im ungarischen Wikipedia-Korpus liefern Kasus-Annotationen, die im deutschen Tagset nicht zur Verfügung stehen, wichtige Hinweise auf die syntaktische Funktion der NP. 58% der ungarischen NPs mit Artikel stehen im linken Feld mit zwei Wörtern im Nominativ, was zwar keine hinreichende Bedingung zur Identifikation darstellt, aber auf eine Bevorzugung der Subjektsfunktion bei dieser Realisierungsform hindeutet. Die übrigen dieser ungarischen NPs sind Träger sog. semantischer Kasus (27%), des Akkusativs (12%) oder Dativs (3%).

Auch in Wiki-PO sind NPs in linken Feldern mit zwei Wörtern der häufigste Phrasentyp, obwohl die polnische Sprache als einzige der sechs Kontrastsprachen keine direkte Entsprechung für den Definit-/Indefinitartikel besitzt. Folglich sind hier durch NPs (7a, b) oder Adjektive erweiterte (8a) nominale Köpfe am häufigsten, wozu auch die adnominalen Verbindungen mit dem adjektivisch deklinierten Demonstrativum *ten* gerechnet werden (8b).

(7a) Nazw-a planetoid-y ...
Name-NOM Planet-GEN
'Der Name des Planeten ...'

(7b) Siedzib-ą gmin-y ...
Sitz-INS Gemeinde-GEN
'Der Gemeindesitz ...'

(8a) Karier-ę piłkarsk-ą ...
Karriere-AKK fußballerisch-AKK
'Die Fußballerkarriere ...'

(8b) Obszar-Ø ten ...
Raum-NOM dieser[NOM]
'Dieser Raum ...'

Interessant sind in diesem Zusammenhang die Verhältnisse im Norwegischen, wo sich die Häufigkeitsverhältnisse von NPs mit Artikel ebenfalls aufgrund sprachspezifischer Struktureigenschaften von denen in anderen Wikipedia-Korpora grundlegend unterscheiden. Definitheit wird im Norwegischen durch Suffi-

gierung direkt am Nomen markiert, was zur Folge hat, dass der NP mit Definitartikel in anderen Sprachen ein einzelnes, definit flektiertes Nomen im Norwegischen entspricht, das auch alleine ein linkes Feld besetzen kann (9). Eine NP mit Indefinitartikel[16] ist im Norwegischen hingegen analog zu ihrem Pendant im Deutschen, Französischen, Italienischen und Ungarischen aufgebaut (10). Im Falle einer definiten, adjektivisch erweiterten NP wird die sog. doppelte Determination angewendet, bei der einem Adjektiv ein Demonstrativum vorausgeht, das die NP formal wie ihre Entsprechung im Deutschen erscheinen lässt (11). Beispiele:

(9) Vulkanen hadde sitt siste utbrudd i 1995.
(http://no.wikipedia.org/wiki/Fogo, Stand 2011)
'Der Vulkan hatte seinen letzten Ausbruch (im Jahr) 1995.'

(10) En vulkan er en geologisk formasjon, [...]
(http://no.wikipedia.org/wiki/Vulkan, Stand 2011)
'Ein Vulkan ist eine geologische Formation, [...]'

(11) Den inaktive vulkanen Aragats (4090 m) er den høyeste toppen.
(http://no.wikipedia.org/wiki/Armenia, Stand 2011)
'Der inaktive Vulkan Aragaz (4090 m) ist der höchste Gipfel.'

Berücksichtigt man die phrasenstrukturellen Besonderheiten des Norwegischen, überrascht es nicht, dass der Anteil des N-Gramms Art+N deutlich niedriger liegt als in den Vergleichssprachen mit pränominalem Definitartikel: Die relative Häufigkeit von das linke Feld besetzenden Nominalphrasen mit Indefinitartikel beträgt in der norwegischen Wikipedia lediglich 2,1%. Im linken Feld mit zwei Wörtern dominieren Nominalphrasen mit Erweiterungsnomen (z.B. Vorname + Nachname) und einfache Präpositionalphrasen.

Im französischen und italienischen Korpus stellen attributiv erweiterte Nominalphrasen nicht nur in linken Feldern mit zwei und drei Wörtern die größte Gruppe, sondern auch in solchen mit vier Wörtern. Es handelt sich um die für die romanischen Sprachen typische Realisierung mit nachgestellter Präpositionalphrase (z.B. *La population du district ...*, *Il nome della Rosa ...*), die funktional häufig dem Genitivattribut im Deutschen entspricht. Genau dieses Pendant, die Nominalphrase mit nachgestellter Nominalphrase im Genitiv (Art+N+Art+N), ist

16 Mit den Flexionsformen *en/ei/et*. Hiervon zu unterscheiden ist das Numerale *én/ett*, das ebenfalls pränominal stehen kann. Beide Verwendungen werden bei Faarlund/Lie/Vannebo (2006: 224f.) zur Determinativ-Subklasse der Quantoren (norw. kvantoren) gezählt und im Korpus unter dem POS-Tag DET kvant zusammengefasst.

auch im deutschen Wikipedia-Korpus der häufigste Realisierungstyp (z.B. *Das Wappen der Stadt* ...) des linken Feldes mit vier Wörtern.

4.2.2 Temporaladverbialia

Temporaladverbialia in verschiedenen Realisierungsformen sind in allen Kontrastkorpora sehr frequente Ausdrücke. Am häufigsten kommen sie als Präpositionalphrasen mit Numerale vor, das in der überwiegenden Zahl der Fälle eine Jahreszahl ist. Das präpositionale Adverbiale mit Jahreszahl ist die typische Realisierungsform im Französischen (*en*+Jahreszahl), Italienischen (*nel*+Jahreszahl) und Norwegischen (*i*+Jahreszahl), z.B.:

(12) I 2001 var dette arbeidet ferdig.
(http://no.wikipedia.org/wiki/Annen_ringvei_(Beijing), Stand 2011)
'(Im Jahr / ?In) 2001 war diese Arbeit fertig.'

Im Polnischen ist sowohl die explizite Angabe der Zeiteinheit *roku* ('Jahr') geläufig, z.B. in der Abfolge Präposition+Jahreszahl+*roku*, als auch dessen Auslassung (*w*+Jahreszahl), wie sie auch in den romanischen Kontrastsprachen und im Norwegischen Usus ist. Das Deutsche gibt hingegen in Präpositionalphrasen entweder die Zeiteinheit an (*Im Jahr(e)*+Jahreszahl) oder realisiert das Adverbiale nur mit der Jahresangabe ohne Präposition (*2001 wurde* ...). In Wiki-PO und Wiki-DE sind präpositionale Adverbialia mit Jahreszahlen die häufigsten dreigliedrigen Ausdrücke im linken Feld entsprechender Größe. Obwohl das korrespondierende Wortklassen-Trigramm P+N+Num in der deutschsprachigen Wikipedia nur an dritter Stelle (12,1%) rangiert, ist seine konkrete Realisierung durch das Wortformen-Trigramm *Im Jahr(e)*+Jahreszahl die häufigste Phrase in einem linken Feld mit drei Wörtern. Das gilt auch für das polnische Pendant *W*+Jahreszahl+*roku*, wobei hier präpositionale Adverbialia mit Jahreszahlen auch auf der Ebene der Wortklassen-Trigramme mit 22% Spitzenreiter bei den relativen Häufigkeiten sind (mit den Stellungsvarianten P+Num+N / P+N+Num). Knapp 80% der P+Num+N-Trigramme bilden eine temporaladverbiale Präpositionalphrase des Typs Präposition+Jahreszahl+*roku*. Auch vergleichbare komplexere Präpositionalphrasen mit vier Wörtern sind häufig, z.B. *Na początku 2012 roku* ... ('Zu Beginn des Jahres 2012 ...' / 'Anfang 2012 ...').

In der deutsch-, italienisch-, norwegisch- und polnischsprachigen Wikipedia zählen punktuelle Datumsangaben mit Tag, Monat und Jahr zu den häufigsten Ausdrücken in linken Feldern mit vier Wörtern. Nur in Wiki-DE handelt es sich hierbei um Präpositionalphrasen mit einem Ordinalzahladjektiv (als konkrete Realisierung des Tetragramms P+Adj+N+Num), die vorwiegend mit einer Ver-

schmelzungsform aus Präposition (mit Dativrektion) und Definitartikel gebildet werden: 95,0% dieser Tetragramme beginnen mit *am* (2,6% mit *zum*), z.B. *Am 12. Dezember 2012* (s. Tabelle 8). Im Französischen, Italienischen und Norwegischen werden Datumsangaben nur mit Definitartikel ohne Präposition, d.h. in Form von Nominalphrasen ausgedrückt. Im Französischen und Italienischen wird der Kalendertag durch eine Kardinalzahl[17] angegeben, in den anderen Kontrastsprachen durch ein Ordinalzahladjektiv (welches aber nur in Wiki-DE und Wiki-NO als Adjektiv und nicht als Numerale klassifiziert ist). Ungarische und polnische Datumsangaben kommen mit drei Wörtern im linken Feld aus. Im Ungarischen wird der Kalendertag als Ordinalzahl[18] ausgedrückt, an die noch das Possessivsuffix der dritten Person *-é-*, das die Zugehörigkeitsrelation zum Monat markiert, und das in Verbindung mit Lokalität stehende Kasussuffix[19] *-n* affigiert wird. Gemäß der ungarischen Schreibkonvention werden nur die letzten beiden Suffixe mit Bindestrich nach der letzten Ziffer ausgeschrieben, z.B. *12-én* (ausbuchstabiert: *tizenkettedikén*). Im Polnischen steht die gesamte Datumsangabe im Genitiv. Die Zugehörigkeitsrelationen zwischen Tag, Monat und Jahr werden innerhalb der komplexen Genitivphrase durch Genitivattribuierung hergestellt. Im Unterschied zu den anderen Vergleichssprachen erscheint auch die Jahreszahl als Ordinalzahladjektiv, das sich auf das weglassbare *roku* bezieht.

Tab: 8: Wortartenklassifikation bei Temporaladverbialia zur Angabe eines Zeitpunkts mit Kalenderdatum (gemäß Wikipedia-Richtlinien)

Korpus	übereinzelsprachliche Wortklassifizierung						
	P	Art	Num	Adj	N	Num	N
DE	*Am*			*12.*	*Dezember*	*2012*	
FR		*Le*	*12*		*décembre*	*2012*	
IT		*Il*	*12*		*dicembre*	*2012*	
NO		*Den*		*12.*	*desember*	*2012*	
PO			*12*		*grudnia*	*2012*	*(rok-u)*
UN			*2012.*		*december*	*12-én*	

17 Nur der erste Tag des Monats wird durch das Ordinalzahladjektiv (Mask.Sg.) frz. *premier* (*1er*) / it. *primo* (*1º*) ausgedrückt.

18 Ungarische Ordinalzahlen werden mithilfe einer Kombination des Suffixes zur Bildung von Bruchzahlen (*-d-*) und dem Suffix *-ik* gebildet.

19 In der ungarischen Grammatikschreibung herrscht in Bezug auf den Status dieses Endsuffixes Uneinigkeit: Pilarský (2013: 187) zählt es zu den Superessivsuffixen. Andere Autoren beschreiben es semantisch weniger spezifisch als Lokativ- oder funktional als Adverbialbildungssuffix.

Die Übersicht in Tabelle 8 zeigt teilweise auch die sprachspezifische Verwendung von Ordinalzeichen in den Wikipedia-Korpora. Als Ordinalzeichen kommen der Punkt (.) sowie hochgestellte Buchstaben, die Flexionsendungen wiedergeben (s. Fußnote 17) vor. Besonders der Punkt wird sehr unterschiedlich verwendet: Im Polnischen wird er im vorliegenden Datumsformat i.d.R. nicht gesetzt, obwohl es sich bei beiden Numeralia um Ordinalia handelt. Im Ungarischen signalisiert ein Punkt nach der Jahreszahl, dass der Tag der Datumsangabe als Ordinalzahl zu lesen ist.

Die häufige Nennung von Kalenderdaten ist ein allgemeines Merkmal der Textsorte Lexikonartikel, wobei Fandrych/Thurmair (2011: 106–108) von einer speziellen Wikipedia-Ausprägung der Textsorte ausgehen, die sich dadurch auszeichnet, dass „einzelne Texteinheiten wesentlich stärker ausgebaut sind" als bei gedruckten Lexikonartikeln (ebd.: 107). Dass in der Online-Enzyklopädie „großzügiger mit Platz umgegangen wird", wie Fandrych/Thurmair (ebd.) allgemein feststellen, wird speziell bei der ausführlichen Angabe von historischen Daten deutlich. Die überdurchschnittlich hohe Frequenz von Numeralia insgesamt kann sprachübergreifend als charakteristisches Merkmal der Wikipedia-Korpora betrachtet werden und stellt in diesem Zusammenhang eine besondere Herausforderung bei der Korpusanalyse dar.

4.2.3 Mehrere Phrasen im linken Feld

Die mehrfache Besetzung des linken Feldes mit Subjekt/Objekt(en) und Adverbialia ist für die untersuchten Kontrastsprachen, mit Ausnahme des Deutschen und Norwegischen, ein typisches Merkmal. In der polnischen Wikipedia liegen bei ca. 20% der P+Num+N-Trigramme temporaladverbiale Präpositionalphrasen mit Jahreszahl ohne Angabe der Zeiteinheit (*roku*) vor, bei denen auf das Temporaladverbiale (P+Num) ein einfaches Nomen als Satzglied (13) folgt. In der französischen Wikipedia folgt häufig ein pronominal realisiertes Subjekt durch ein Komma getrennt auf eine Nominal- oder Präpositionalphrase in temporaladverbialer Funktion (14). Bei über 97% der französischen P+Num+Pro-Trigramme ist das Numerale eine Jahreszahl. In den Sprachen, in denen solche Mehrfachbesetzungen des linken Feldes möglich sind, können auch gleich mehrere Adverbialia vorkommen. Im ungarischen Beispiel (15) vom Typ Num+N+N teilt sich das Subjekt das linke Feld mit zwei Adverbialia.

(13) Od 1926 kościół-Ø jest katedr-ą.
Seit 1926 Kirche-NOM ist Kathedrale-INS
(http://pl.wikipedia.org/wiki/Telsze, Stand 2011)
'Seit 1926 ist die Kirche eine Kathedrale.'

(14) En 2002, il a reçu le grammy award.
In 2002 er hat erhalten[PTCP] den grammy award.
(http://fr.wikipedia.org/wiki/David_Darling, Stand 2011)
'2002 hat er den Grammy Award erhalten.'

(15) 1968-ban Jung Kaliforniá-ba költözö-tt barát-já-val,
1968-INE Jung Kalifornien-ILL zieh-PST.3SG Freund-POSS.3SG-COM
Tuná-val.
Tuna-COM
(http://hu.wikipedia.org/wiki/George_Jung, Stand 2011)
'1968 zog Jung mit seinem Freund Tuna nach Kalifornien.'

Da die syntaktische Einbettungstiefe und Funktion der Phrasen in den Wikipedia-Korpora nicht annotiert ist, wurden in Bezug auf die Mehrfachbesetzungen von linken Feldern keine entsprechenden quantitativen Analysen durchgeführt. Die Frequenzbestimmung von Wortklassen-N-Grammen alleine liefert noch keine hinreichenden syntaktischen Informationen: Beispielsweise sind in der ungarischen Wikipedia Art+N+Adv[20] und Art+N+N die frequentesten Trigramme. Auch Art+N+N kann mit zwei eigenständigen Phrasen besetzt sein, z.B. wenn das zweite N durch ein Adverbiale in Form eines Nomens mit semantischem Kasussuffix (z.B. *Berlin-ben* 'in Berlin') realisiert wird, oder wenn das erste N als Attribut zum zweiten N fungiert und diese Zugehörigkeitsrelation durch ein Possessivsuffix am zweiten N ausgedrückt wird (z.B. *A szezon végén* ... 'am Ende der Saison ...'). Nomina/Nominalphrasen können bei Mehrfachbesetzung im linken Feld aber auch als Objekt fungieren, d.h. Subjekt und Objekt können gemeinsam im linken Feld erscheinen, wie der folgende ungarische (vgl. Uzonyi/Daböczi in diesem Band) und der italienische Beleg für das Tetragramm Art+N+Art+N illustrieren:

(16) A minisztérium-Ø a részletfizetés-t engedélyez-t-e, [...]
Das Ministerium-NOM die Ratenzahlung-AKK genehmig-PST-OBJ.3SG
(http://hu.wikipedia.org/wiki/Baja_Ferenc, Stand 2011)
'Das Ministerium genehmigte die Ratenzahlung, [...] '

20 Zu den häufigsten Adverbien, die hier auf die Nominalphrase folgen, zählt u.a. der Negator *nem*, z.B.:
A döntés-Ø nem volt-Ø egyszerű.
die Entscheidung-NOM nicht war-PST.3SG einfach
(http://hu.wikipedia.org/wiki/FC_Bayern_München, Stand 2011).
'Die Entscheidung war nicht einfach.'

(17) Le vaccinazion-i, gli affidatar-i sono ten-ut-i, [...] a provved-ere
Die Impfung-PL die Pflegeeltern-PL sind halt-PTCP-PL zu sorg-INF
(http://it.wikipedia.org/wiki/Affido_familiare, Stand 2011)
'Die Pflegeeltern sind gehalten, für die Impfungen zu sorgen, [...]'

Von den untersuchten Kontrastsprachen weisen das Deutsche und das Norwegische bestimmte Restriktionen in Bezug auf die Mehrfachbesetzung des linken Feldes auf (vgl. Bassola/Schwinn in diesem Band). Zwischen den Besetzungsregularitäten des deutschen Vorfeldes und seinem norwegischen Pendant, dem *forfelt*[21], existiert eine weitreichende Parallelität. Für die skandinavischen Sprachen gilt die Stellungsfelderanalyse von Diderichsen (1941–1942) als grundlegend[22], in dessen Tradition auch die norwegische Referenzgrammatik (Faarlund/Lie/Vannebo 2006) steht.

4.2.4 Pronomina

Die Kontrastsprachen Italienisch, Polnisch und Ungarisch zählen zu den Pro-drop-Sprachen, das Deutsche, Französische und Norwegische hingegen nicht (Pronomen-Sprachen). Der Wegfall von Subjektspronomina (im Ungarischen betrifft dies unter bestimmten Voraussetzungen auch die Objektspronomina) wirkt sich zwangsläufig auf die relativen Häufigkeiten einzelner Wortklassen in den linken Feldern aus, da diese in allen sechs Kontrastsprachen der topologisch prädestinierte Ort für das Vorkommen von Subjektspronomina sind: Der Anteil von Personalpronomina in den linken Feldern der Pro-drop-Sprachen ist folglich geringer als in den anderen Kontrastsprachen (s. Abbildung 3). In fast allen Kontrastkorpora ist die relative Häufigkeit von Personalpronomina im linken Feld mit nur einem Wort am größten. Am geringsten ist sie im linken Feld mit genau zwei

21 Faarlund/Lie/Vannebo (2006) teilen den norwegischen Satz ausgehend von der Position der finiten und infiniten Verbformen in *forfelt*, *midtfelt* und *sluttfelt* ein, wobei – im Unterschied zur deutschen Satzklammer – die Verbformen als Bestandteile der Stellungsfelder betrachtet werden: Das *midtfelt* wird vom finiten, das *sluttfelt* von den infiniten Verbformen eröffnet. Die Parallelität in Bezug auf Besetzungsregularitäten gilt im Übrigen auch für die Stellungseinheiten links des finiten Verbs, die nicht als Vorfeldeinheiten betrachtet werden.

22 Diderichsens Modell ähnelt dem von Drach (1937) (vgl. 3.1), ohne sich aber explizit auf dieses zu beziehen. Eine gemeinsame wissenschaftshistorische Basis der Modelle erscheint vor dem Hintergrund der ähnlichen Terminologie und nahen Verwandtschaft der Sprachen plausibel (Andersen 2012: 37). Diderichsen gibt allerdings rückblickend an, Drachs „Grundgedanken der deutschen Satzlehre" von 1937 erst später kennengelernt zu haben (Diderichsen 1976: 323).

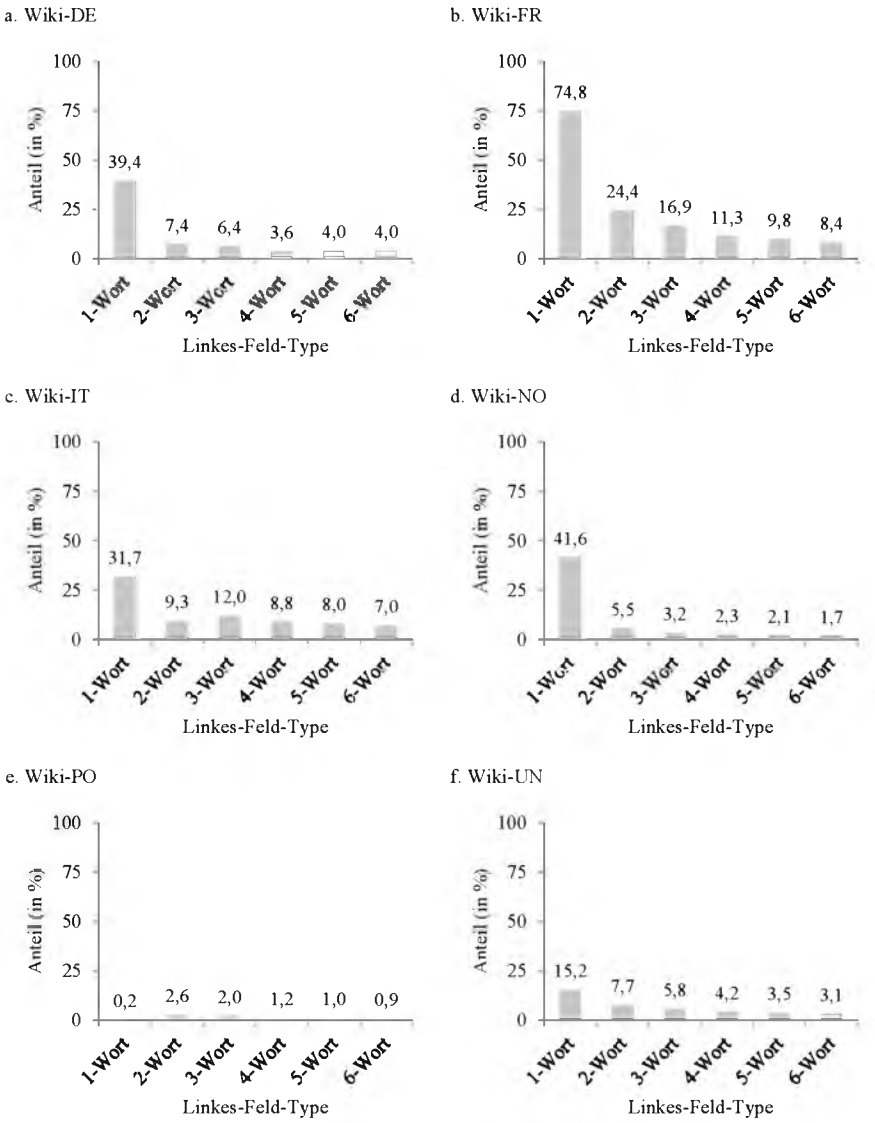
Wörtern, was sich dadurch erklärt, dass Pronominalphrasen im Vergleich zu Nominalphrasen nur wenige Erweiterungsmöglichkeiten besitzen und keine Artikel bei sich führen. In dieser Position dominieren überproportional stark Art+N-Bigramme. Im Polnischen als Nicht-Artikel-Sprache bleibt dieser Effekt aus. Wiki-PO besitzt im linken Feld mit einem Wort den höchsten prozentualen Anteil von Nomina (65,4%) und gleichzeitig den geringsten Anteil an Pronominalphrasen (0,2%). Die leicht erhöhte Häufigkeit von Pronomina im linken Feld mit zwei Wörtern ist auf den Umstand zurückzuführen, dass in diesem Korpus auch die adnominalen Possessiva, die sich im Polnischen morphologisch nicht von den selbstständigen unterscheiden, zur Klasse der Personalpronomina gezählt werden. Bei 93% der Pronomen-Vorkommen handelt es sich um Pro+N-Bigramme, z.B. *Jego brat* ... 'Mein Bruder ...'.

In linken Feldern mit mehr als zwei Wörtern steigt der prozentuale Anteil von Personalpronomina in allen Kontrastsprachen wieder leicht an, da sie vermehrt in Mehrfachbesetzungen des linken Feldes vorkommen.

Im italienischen Korpus wirkt sich nicht nur pro drop auf die relative Häufigkeit von Personalpronomina aus, sondern auch die Pronomen-Enklise. Unbetonte Objektspronomina (z.B. *la*, *lo*, *mi*, *si* usw.) und Pronominaladverbien (*ci*, *ne*, *vi*) werden u.a. an nicht-finite Verbformen wie Imperativ, Infinitiv (18c), Gerundium (18b), Partizip Perfekt und andere Objektformen (18a) angehängt und dann zusammen geschrieben (z.B. *dirglielo* 'es ihm (zu) sagen'). Solche enklitischen Verbindungen wurden bei der morphosyntaktischen Annotation aber nicht weiter segmentiert, es wurde lediglich die Wortklasse der Basis bei der Vergabe der POS-Tags berücksichtigt. Mit anderen Worten: Eine Wortform kann bei anderer Zählung bis zu drei Wortformen (Basis + ein bzw. zwei Enklitika) entsprechen (vgl. (18c)). Ein linkes Feld mit nur einem Wort kann mehrfach pronominal besetzt sein und das selbst dann, wenn das Subjektspronomen unausgedrückt bleibt (18a). Eine manuelle Zählung hat ergeben, dass 5,5% der nicht-finiten Verbformen (ohne Imperative) in den untersuchten linken Feldern enklitische Pronomina bei sich führen.

(18) a.	Glie=lo	spieg-o.	b.	Ved-endo=l-a,...	c. ricordar=se=l-i
	ihm=es	erklär-1SG		seh-GER=sie-F.SG	erinner=sich=sie-M.PL
	'Ich erkläre es ihm.'			'Sie sehend, ...'	'sich an sie erinnern'

Abb. 3: Prozentualer Anteil von Personalpronomina in linken Feldern in den Wikipedia-Korpora



Der für eine Pro-drop-Sprache überraschend hohe Anteil von Personalpronomina (17%) im linken Feld mit nur einem Wort geht vorwiegend auf einen besonderen Konstruktionstyp des Italienischen zurück, der mit dem Reflexivum *si* gebildet wird. Dies zeigt sich bei einer genaueren Betrachtung der linken Felder mit einem Wort: Personalpronomina, die ausschließlich Subjekte realisieren können (*io*, *tu*, *egli*, *ella*), machen 2,1% aus, Objektspronomina 1,4% und Personalpronomina, deren syntaktische Funktion anhand der morphologischen Form nicht eindeutig bestimmt werden kann, kommen auf 3,5%. Die Reflexivkonstruktion mit *Si* (+ finites Verb in 3. Pers. Sg.) kommt in 10% der Fälle vor. Es handelt sich dabei um einen sehr gebräuchlichen Konstruktionstyp des Italienischen, bei dem das Reflexivum *si*, das nicht wegfallen kann, besonders häufig das linke Feld alleine besetzt. Die reflexivierten Verben haben dabei unterschiedliche Lesarten, die Schwarze (1995: 182ff.) abstrahierend-generisch[23] (19, 20) bzw. rückbezüglich (21) nennt, z.B.:

(19) Si viv-e solo due volt-e.
sich leb-3SG nur zwei mal-PL.
'Man lebt nur zweimal.' (Filmtitel)

(20) Si cambi-ano nom-e e color-i [...]
sich änder-3PL Name-SG und Farbe-PL
'Name und Farben wurden geändert ...' (https://it.wikipedia.org/wiki/Grassina, Stand 2011)

(21) Si cambi-ano, [...] e finalmente si mett-ono a dormire.
sich umzieh-3PL und endlich sich leg-3PL zu schlafen
'Sie ziehen sich um [...] und legen sich endlich schlafen.'
(http://it.wikipedia.org/wiki/Concerto_di_violoncello, Stand 2011)

Ohne das Reflexivpronomen sinkt der Anteil von Pronominalphrasen in dieser Position auf 15,5% und erreicht damit das Frequenz-Niveau von Pronominalphrasen in der ungarischen Wikipedia (15,2%). Insgesamt macht das Italienische viel häufiger Gebrauch von Objektspronomen und von Pronominaladverbien (die in

23 Es kann sich bei intransitiven Verben wie in (19) um eine Ersatzkonstruktion für ein im It. nicht vorhandenes generalisierendes Subjektspronomen handeln (it. *si* impersonale, vgl. dt. *man*, frz. *on*) oder bei transitiven Verben wie in (20) um eine passiv-ähnliche Konstruktion in der 3. Pers. Sg./Pl., bei der das Objekt des aktiven Verbs als Subjekt erscheint (it. *si* passivante, s. Schwarze 1995: 187). Betrachtet man wie Schwarze diese Reflexivierung „als Verfahrung der Reduktion von Valenzen" (ebd.: 182), liegt in (19, 20) Subjekt-Reduktion, in (21) Objekt-Reduktion in Verbindung mit Wegfall des Subjektspronomens vor.

Wiki-IT als Pronomina annotiert sind) als das Polnische. Insgesamt ist im italienisch- und ungarischsprachigen Korpus der Pronominalphrasen-Anteil im linken Feld mit nur einem Wort erwartungsgemäß deutlich geringer als bei den Pronomen-Sprachen.

4.2.5 Textsortenspezifizität

Während bereits an verschiedenen Stellen auf spezifische Merkmale der Textsorte Wikipedia-Artikel eingegangen wurde, steht eine direkte korpusanalytische Kontrastierung der Daten in diesem Zusammenhang noch aus. Eine solche Kontrastierung kann im Rahmen der vorliegenden Studie nur punktuell und exemplarisch vorgenommen werden. Ausgangspunkt ist hierbei ein Vergleich des deutschsprachigen Wikipedia-Korpus mit anderen deutschsprachigen, morphosyntaktisch annotierten Korpora (in COSMAS II), die jeweils eine relativ homogene Textsortenstruktur aufweisen. Die Häufigkeitsverteilung von Worttokens in Wiki-DE wurde mit jener in folgenden Korpora verglichen, die unterschiedliche sprachliche Domänen abdecken:

Tab. 9: Deutschsprachige Vergleichskorpora

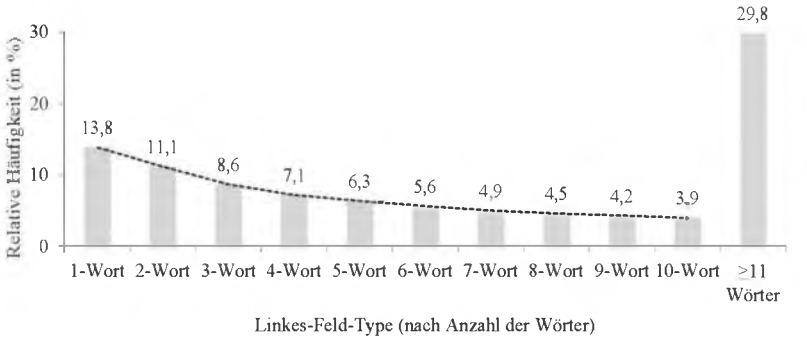
Korpus	Inhalt	Größe (MW)
Thomas Mann	60% Romane und Erzählungen 34% Reden und Aufsätze 6% Nachträge zur Gesamtausgabe (1893–1955)	3,47
Mannheimer Morgen	Mannheimer Morgen (2006–2009)	85,13
Die Zeit	Die Zeit (2006–2009)	31,62
Reden und Interviews	Reden der Bundestagsfraktion Bündnis 90/DIE GRÜNEN (2002–2006)	1,93
Biografische Literatur	90% Victor Klemperer: Tagebücher, Notizen eines Philologen (1918–1959) 10% Alfred Kerr: Berliner Briefe (1895–1900)	1,90
Wikipedia (DE)	WPD11 Wikipedia.de Artikel (2011)	551,09

Das Thomas-Mann-Korpus wurde ausgewählt, weil es einen hohen Anteil literarischer Texte enthält. Eine speziellere Form literarischer Texte repräsentieren die Tagebücher in „Biografische Literatur“. Beide Korpora sind weitgehend autorenspezifisch. Mit dem „Mannheimer Morgen“ (regionale Tageszeitung) und „Die Zeit“ (überregionale Wochenzeitung) sind Pressetexte unterschiedlichen Typs

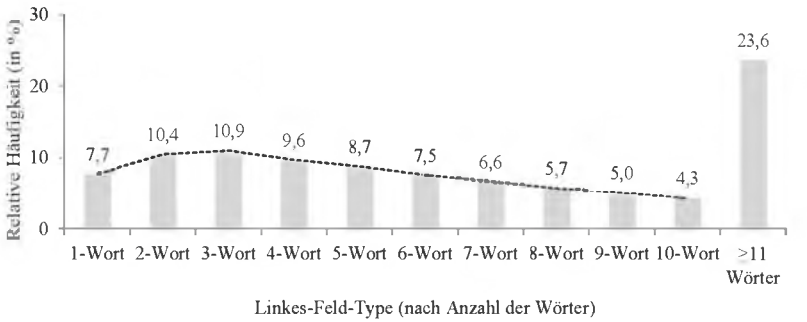
vorhanden. Das Korpus „Reden und Interviews“ beinhaltet politische Texte (Bundestagsreden).

Abb. 4: Relative Häufigkeit von Worttokens pro Linkes-Feld-Type in deutschsprachigen Korpora

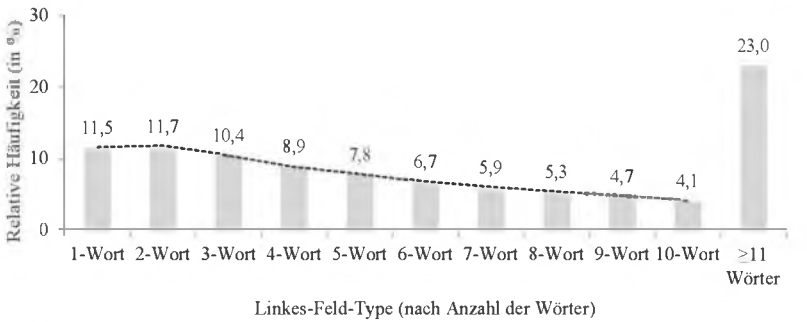
a. Thomas Mann



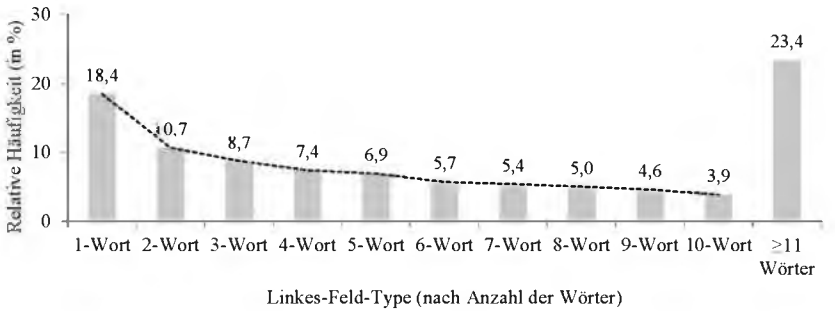
b. Mannheimer Morgen



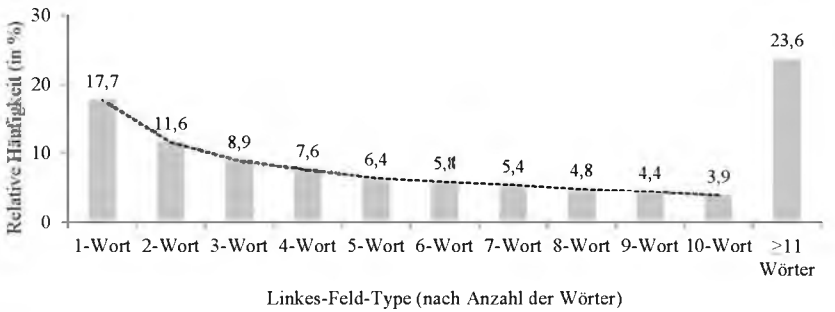
c. Die Zeit



d. Reden und Interviews



e. Biografische Literatur



f. Wikipedia (DE)

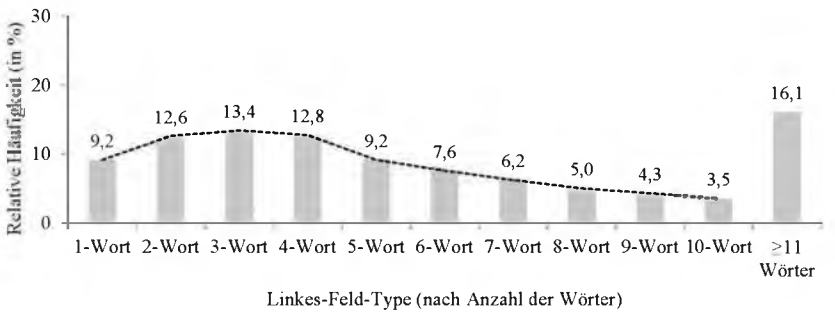


Abbildung 4 zeigt die relative Häufigkeit von Worttokens in den linken Feldern der Vergleichskorpora aus Tabelle 9. Ausgehend von der Annahme, dass ähnliche Kurvenverläufe auf texttypologische Gemeinsamkeiten hindeuten, lassen sich zwei Korpus-Gruppen ausmachen: Drei Korpora zeichnen sich durch ein Maximum der Worttoken-Frequenz beim linken Feld mit zwei Wörtern (Die Zeit) bzw. drei Wörtern (Wikipedia, Mannheimer Morgen) aus. Die Worttoken-Frequenzen in den übrigen Korpora weisen keine charakteristische Glockenkurven auf, da hier Worttokens von linken Feldern mit nur einem Wort mit großem Abstand am häufigsten vorkommen (Thomas-Mann-Korpus, Biografische Literatur, Reden und Interviews).

Anhand von sprachlichen Merkmalen, die in den Vergleichskorpora jeweils unterschiedlich ausgeprägt sind, kann korpusanalytisch nachgewiesen werden, dass texttypologische Kriterien den Kurvenverlauf der Worttoken-Frequenz maßgeblich mitbestimmen. Bei der Datenanalyse erwiesen sich Personalpronomina (vgl. 4.2.4) als besonders gut geeignete Indikatoren, um die Textsorten- bzw. Korpusspezifizität von sprachlichen Merkmalen aufzuzeigen: Die in Tabelle 10 wiedergegebene relative Häufigkeit von (irreflexiven) Personalpronomina in linken Feldern mit genau einem Wort korreliert exakt mit der Einteilung der Kontrastkorpora in zwei distinkte Gruppen mit charakteristischer Worttoken-Frequenz (vgl. Abbildung 4). In der Wikipedia/Zeitungstexte-Gruppe liegt der Anteil von Personalpronomina unter 30%, in der Gruppe, die Reden, Tagebücher und literarische Texte umfasst, liegt der Anteil bei ca. 50%. Diese strukturellen Gemeinsamkeiten lassen den Schluss zu, dass bei Korpora, deren Texte sich durch die Bevorzugung von pronominaler Referenz auszeichnen, die Häufigkeitsverteilung von Worttokens maßgeblich durch die relative Kürze von Pronominalphrasen bestimmt ist.

Tab. 10: Irreflexive Personalpronomina in linken Feldern mit nur einem Wort

Korpus	Rel. Häufigkeit von Personalpronomina (in %)
Mannheimer Morgen	17,2
Wikipedia (DE)	25,2[24]
Die Zeit	28,4
Thomas Mann	47,7
Reden und Interviews	49,6
Biografische Literatur	54,2

24 Der Wert weicht leicht vom in Tabelle 6 angegebenen Stichproben-Wert ab, da hier keine manuelle Überprüfung der Annotationen durchgeführt wurde.

Gerade die ausgeprägten Unterschiede bei der Häufigkeit von Personalpronomina zwischen den untersuchten Korpora derselben Sprache macht deutlich, wie stark sich neben dem Einfluss sprachspezifischer Merkmale insbesondere auch textsortenspezifische Struktureigenschaften auf die quantitativen Verhältnisse in den linken Feldern der Wikipedia-Korpora auswirken.

5 Fazit

Die Resultate der kontrastiven Korpusanalysen lassen die Wikipedia-Korpora wenig geeignet erscheinen, als „universelle" Korpora zu fungieren, die sämtliche sprachspezifischen Merkmale repräsentativ abzubilden vermögen. Diesen Erwartungen, die an entsprechende Referenz- und Nationalkorpora (sofern vorhanden) gerichtet werden, können die Texte der Online-Enzyklopädie trotz ihrer thematischen Breite und ihres beträchtlichen Umfangs nicht in gleicher Weise gerecht werden. Sprachliche Strukturmerkmale können nur unter Berücksichtigung der Textsortenspezifika, die in dieser Studie qualitativ und quantitativ partiell und exemplarisch herausgearbeitet wurden, adäquat erfasst werden. In diesem Kontext verfügen die vorliegenden (und zukünftigen) Wikipedia-Korpora – im Rahmen der durch die Textsorte gesetzten Grenzen – über ein enormes Analysepotenzial für kontrastive Fragestellungen und stellen somit eine wertvolle Ressource für die Korpuslinguistik dar.

Abkürzungen der Interlinearglossierung

AKK	Akkusativ	M	Maskulinum
COM	Komitativ	NOM	Nominativ
F	Femininum	OBJ	objektive Konjugation
GEN	Genitiv	POSS	Possessum
GER	Gerundium	PL	Plural
ILL	Illativ	PST	Vergangenheitsform
INE	Inessiv	PTCP	Partizip Perfekt
INF	Infinitiv	SG	Singular
INS	Instrumental		

Literatur

Altmann, Hans (1987): Zur Problematik der Konstitution von Satzmodi als Formtypen. In: Meibauer, Jörg (Hg.): Satzmodus zwischen Grammatik und Pragmatik. Referate anläßlich der 8. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Heidelberg 1986. (= Linguistische Arbeiten 180). Tübingen: Niemeyer. 22–56.

Andersen, Christiane (2012): Wortfolge im gesprochenen Deutsch. Markiertheit vs. Unmarkiertheit als Kriterien der Nachfeldbesetzung aus kontrastiver Perspektive. In: Zielsprache Deutsch 39, 1. 35–58.

Brockhaus-Enzyklopädie (2006): 21., völlig neu bearb. Aufl. Leipzig/Mannheim: Brockhaus.

Bubenhofer, Noah/Haupt, Stefanie/Schwinn, Horst (2011): A comparable Wikipedia corpus: From Wiki Syntax to POS Tagged XML. *Arbeiten zur Mehrsprachigkeit* (Folge B Nr. 96). 141–144.

Diderichsen, Paul (1941-1942): Sætningsbygningen i Skånske Lov. Fremstillet som Grundlag for en rationel dansk Syntaks. In: Acta Philologica Scandinavica 15. 1–252.

Diderichsen, Paul (1976): Die Satzglieder und ihre Stellung – Nach dreißig Jahren. In: Diderichsen, Paul: Ganzheit und Struktur. Ausgewählte sprachwissenschaftliche Abhandlungen. (= Internationale Bibliothek für allgemeine Linguistik 30). München: Fink. 320–344.

Drach, Erich (1937): Grundgedanken der deutschen Satzlehre. Frankfurt a.M.: Diesterweg.

Faarlund, Jan Terje/Lie, Svein/Vannebo, Kjell Ivar (2006): Norsk referansegrammatikk. 4. Aufl. Oslo: Universitetsforl.

Fandrych, Christian/Thurmair, Maria (2011): Textsorten im Deutschen. Linguistische Analysen aus sprachdidaktischer Sicht. (= Stauffenburg Linguistik 57). Tübingen: Stauffenburg.

Hoffmann, Lothar (1988): Vom Fachwort zum Fachtext. Beiträge zur angewandten Linguistik. (= Forum für Fachsprachen-Forschung 5). Tübingen: Narr.

Pilarský, Jiří (2013): Deutsch-ungarische kontrastive Grammatik. (= Veröffentlichungen des Instituts für Germanistik an der Universität Debrecen. Studienmaterialien 10). Debrecen: Debreceni Egyetemi Kiado.

Przepiórkowski, Adam/Woliński, Marcin (2003): A flexemic tagset for Polish. In: Proceedings of the workshop on morphological processing of Slavic languages, EACL 2003. Budapest. 33–40.

Schwarze, Christoph (1995): Grammatik der italienischen Sprache. 2. Aufl. Tübingen: Niemeyer.

Wikipedia (2014a): Wikipedia: Allgemeinverständlichkeit. http://de.wikipedia.org/w/index.php?title=Wikipedia: Allgemeinverst%C3%A4ndlichkeit&oldid=132451177 (Stand 2014).

Wikipedia (2014b): Wikipedia: Datumskonventionen. http://de.wikipedia.org/w/index.php?title=Wikipedia:Datumskonventionen&oldid=128578053 (Stand 2014).

Wikipedia (2014c): Wikipedia: Wie schreibe ich gute Artikel. http://de.wikipedia.org/w/index.php?title=Wikipedia:Wie_schreibe_ich_gute_Artikel&oldid=131001378 (Stand 2014).

Wikipedia (2014d): Wikipêdia: Formai útmutató. http://hu.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Formai_%C3%BAtmutat%C3%B3&oldid=14642511 (Stand 2014).

Wikipedia (2014e): Pomoc: Styl – poradnik dla autorów. http://pl.wikipedia.org/w/index.php?title=Pomoc: Styl_%E2%80%93_poradnik_dla_autor%C3%B3w&oldid=40443996 (Stand 2014).

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7). Berlin/New York: de Gruyter.

Zimmer, Dieter E. (2007): Gutes Deutsch. In: Burkhardt, Armin (Hg.): Was ist gutes Deutsch? Studien und Meinungen zum gepflegten Sprachgebrauch. (= Thema Deutsch 8). Mannheim: Dudenverlag. 381–392.